CAMBRIDGE LAW REVIEW

VOLUME III
2018

Editor-In-Chief
Yen Jean Wee

Proudly Supported By
Cambridge University Law Society

CAMBRIDGE LAW REVIEW

VOLUME III
2018

Editor-In-Chief
Yen Jean Wee

Proudly Supported By
Cambridge University Law Society

Cambridge Law Review

Editors

Agnes So*
Alessandra Moroni
Alexander Fong
Alice Nagel
Alyssa P'ng
Andrea Chong
Andrea Trozzo
Ashley James Hannay
Bridget Yim Choi Wai
Danielle Carrington*
Eamonn Butler
Elizabeth Huang
Emmanuel Alabi
Gordon Chung Chun Wai
Helen Jennings*
Hend Hanafy
Jamie Saleh
Jasmine On Yan Chow
Jonathan Fung
Laura Spiteri
Marios Eleftherios Sarantopoulos
Matthew Isaacs*
Matthew Keliris-Thomas*
Michael Foran
Natalie Yeo
Neli Frost
Ntayi Anfani Bandawa
Orestis Sherman*
Raveena Sethia
Rebecca Fennessy
Roderick Freeman
Shaun Matos
Siang Yee Chua
Stephanie Chng
Sven-Hendrik Schulze
Teh Wen Jin
Theano Karanikioti
Thomas Papadogiannis Varouchakis
Timothy Ong Kah Yong
Tomas Strelecek
Yan Lai
Zachary Seow

Website Design

Jared Kang

* Indicates that an editor was not involved in the substantive and technical editorial process.

Editors' Introduction to the Third Volume of the Cambridge Law Review

It is with great pleasure that I present the third Volume of the *Cambridge Law Review*. It has been a busy year for the journal. We welcomed our new Honorary Board—comprising many distinguished former judges at both the domestic and international levels, academics, and practising barristers—from whose advice and guidance the journal has benefited tremendously. This year also saw the launch of our supplementary journal, *De Lege Ferenda*, helmed by a separate team of student editors and specifically targeted at undergraduate students publishing their work for the first time. In addition, we have deepened our collaboration with the Cambridge University Law Society, working together to organise a guest lecture by Professor Andrew Murray of the London School of Economics on the use of algorithms and profiling in counter-terrorism.

This Volume seeks to offer insight into a range of contemporary legal issues, from the implications of new data technologies for human rights and the regulation of cyberspace, to the legal protection afforded to medical-humanitarian non-governmental organisations in situations of armed conflict, to police deception and the 2017 case of *United States v Spivey*. The international and comparative perspective adopted by many of the articles also serves as a reminder of how much legal discourse can gain from an understanding of the law and practice of other jurisdictions. Beyond this, however, this Volume also takes a step back to tackle some broader questions, such as the role of emotion in legal decision-making and the meaning and purpose of the separation of powers. I hope that the breadth and depth of the articles we have selected will be of interest to British and international, student and professional audiences alike, and we are proud to publish such thoughtful and incisive pieces of legal scholarship.

I owe heartfelt thanks to our team of undergraduate and postgraduate student editors, and above all to the Vice Editors-in-Chief—Jared Kang, Eirini Kikarea, and Ritwika Sharma—for their tireless work and dedication to ensuring the highest standards of quality for this Volume of the journal. I would also like to thank the Cambridge University Law Society for their continued support, and Craig Slade at Crucible Creative. I wish the incoming Editorial and Managing Boards every success with the fourth volume, and I look forward to the future growth of the *Cambridge Law Review* with confidence that they will take the journal to new heights.

Yen Jean Wee

Table of Contents

Typeset in Baskerville
Printed & Bound in the United Kingdom
for the Cambridge Law Review
by Crucible Creative

Jekyll and Hyde Creditors: The Strange Case of Future and Contingent Debts

Paulina Fishman*

I. Introduction

Normally, a company is under no obligation to settle its debts *before* they become due and payable. Two obvious exceptions exist: one, when the company is being wound up; or two, when it wishes to embark on a corporate restructuring scheme or arrangement that includes all of its creditors. In an English winding up, with limited exceptions,[1] "[a]ll claims by creditors… are provable as debts against the company… whether they are present or future, certain or contingent, ascertained or sounding only in damages".[2] Australia has much in common with England in this area of the law.[3] Also subject to certain exceptions, in an Australian winding up, "all debts payable by, and all claims against, the company (present or future, certain or contingent, ascertained or sounding only in damages), being debts or claims the circumstances giving rise to which occurred before the relevant date, are admissible to proof against the company".[4] Thus, a company in either jurisdiction may find itself paying creditors whose claims have not matured—either when it is going into liquidation, in which case it *must* deal with them, or when it is restructuring, in which case it *may* deal with them. The purpose of this article is to

* B.Comm./LL.B. (Hon) (Monash), M.St. in Legal Research (Oxon). The author would like to thank Associate Professor Kristin van Zwieten, University of Oxford, for her guidance and advice in supervising the dissertation on which this article is based. The author is also grateful to her examiners and reviewers for their constructive feedback.

1 See Insolvency (England and Wales) Rules 2016, SI 2016/1024, r 14.2(2)–(5).

2 ibid r 14.2(1). See also r 14.1(5).

3 See, for example, Re HIH Casualty and General Insurance Ltd [2008] UKHL 21, [2008] 3 All ER 869 [2] (Lord Hoffman); Re HIH Casualty and General Insurance Ltd [2005] EWHC 2125 (Ch), [2006] 2 All ER 671 [34] (David Richards J).

4 Corporations Act 2001 (Cth), s 553(1).

explore the role that future creditors, and more importantly, contingent creditors, play in these processes in England and Australia.

Part II of the article begins by briefly defining future claims (held by future creditors), and then dwells at length on the more complicated concept of contingent claims (held by contingent creditors). In delineating the latter, it emerges that the 'contingent claim' is a slightly narrower concept in Australia than in England. Thus, the scope of the problem identified later in the article is not entirely the same in the two jurisdictions. Next, Part III examines the implications of future and contingent creditors participating in liquidations and reorganisation regimes. It is demonstrated that their claims inevitably create a 'Jekyll and Hyde' problem: the value of these claims will either be underestimated or overestimated, though which it will be is unknown. This, in turn, results in unfair distributions and unfair voting, discussed in Part IV.[5] The article concludes with a discussion of how the Jekyll and Hyde problem may be reduced or ameliorated in England and Australia.

II. Future and Contingent Claims

A. Existing legal obligation

If a company owes a debt which is already due and payable, it is a present debt. By contrast, if a company must pay a sum of money in the future, it owes a future debt, though the creditor is not yet entitled to payment. Contingent debts involve a further layer of complexity: these are debts which may or may not become due in the future. In the Australian High Court case of *Community Development Pty Ltd v Engwirda Construction Company*,[6] Kitto J (with whom Barwick CJ and Windeyer J agreed)[7] authoritatively held that, for future and contingent debts, "there must be an *existing obligation* and that out of that obligation a liability on the part of the company to pay a sum of money will arise in a future event, whether it be an event that *must* happen or only an event that *may* happen".[8] If the event *must* occur, it is a future claim; if the event *may* occur, it is a contingent claim.[9] For example, the respondent in *Engwirda* was held to be a contingent creditor because "the appellant was… under a contractual obligation to pay to the respondent the amount, if any,

5 The author has previously written about this problem of valuing debts that are not due and payable, and the unfairness that may result: see Paulina Fishman, 'Statutory Misinterpretation: Rash Holding in Brash Holdings' (2017) 45 Federal Law Review 199, 210–211; Paulina Fishman, 'Voluntary Arrangements and the "Clean Slate" Mess' (2018) 29 Journal of Banking and Finance Law and Practice 109, 112–114.

6 *Community Development Pty Ltd v Engwirda Construction Company* (1969) 120 CLR 455.

7 ibid 457 (Barwick CJ), 460 (Windeyer J).

8 ibid 459 (emphasis added), discussing a quote from *Re William Hockley Ltd* [1962] 2 All ER 111 (Ch) 113 (Pennycuick J). cf *Re SBA Properties Ltd* [1967] 2 All ER 615 (Ch), 618 (Pennycuick J).

9 See especially *Expile Pty Ltd v Jabb's Excavations Pty Ltd* (2004) 22 ACLC 667 [37] (Palmer J).

which *might* be found by an arbitrator to be due to it under the building contract".[10] Thus, a contingent claim in Australia requires an existing legal obligation to pay in certain circumstances.[11]

The English understanding of future claims seems to be the same as in Australia. In *Re Liberty International plc*, Norris J succinctly described future claims as "claims not presently due but which are certain to accrue due in the future".[12] English courts have also held that "[t]he essence of a contingent liability… [is] that… the event on which it depends may never happen".[13] Some jurists, however, believe that "[i]n England… an existing liability is not an essential element of a contingent debt".[14] It is respectfully submitted that this is a misconception stemming from an overly literalist reading of the seminal case of *Winter v IRC*. It is true that Lord Reid in *Winter* rejected the argument "that there must be an existing obligation"[15] for a contingent liability to exist. Taken as a whole, however, Lord Reid's speech reveals that this rejection was premised on the understanding of an existing obligation as something presently operative (rather than dormant). Lord Reid insisted that a contingent debt requires an extant *commitment* to pay in certain circumstances,[16] and emphasised that the concept "only include[s] liabilities which in law *must arise* if one or more things happen".[17] Lord Birkett stressed that a contingent debtor must be "*under a liability* to pay… in the circumstances provided".[18] Lord Guest likewise held that a contingent debtor must be "*automatically involved by the operation of law* in the payment… in the circumstances defined".[19] Thus, in substance—though not in form—*Winter* is authority for the proposition that a contingent liability in England must be underpinned by an existing legal obligation.

In the minority in *Winter*, Lord Hodson (with whom Lord Tucker agreed)[20] expressly held that "[t]here can be no true contingent liability unless there is an *existing legal obligation* under which a payment will become due on the happening of a future unascertained event or events".[21] As explained above, however, this

[10] *Engwirda* (n 6) 461–462 (Owen J) (emphasis added). See also 459–460 (Kitto J).

[11] See, for example, *Australian Gypsum Industries Pty Ltd v Dalesun Holdings Pty Ltd* (2015) 197 FLR 1 [70]– [73] (Buss JA); *BE Australia WD Pty Ltd (subject to a Deed of Company Arrangement) v Sutton (2011) 82 NSWLR 336* [105], [107] (Campbell JA, McColl and Young JJA agreeing) (Sutton); *McLellan v Australian Stock Exchange Ltd* (2005) 144 FCR 327 [16].

[12] *Re Liberty International plc* [2010] EWHC 1060 (Ch), [2010] 2 BCLC 665 [15].

[13] *Winter v IRC* [1963] AC 235 (HL) 251 (Lord Reid) (emphasis added). See also 249.

[14] *McLellan* (n 11) [9] (Finkelstein J). See also [16]; *Glenister v Rowe* [2000] Ch 76 (CA) 84.

[15] *Winter v IRC* (n 13) 249. See also 251 (Lord Reid), 253 (Lord Birkett), 263 (Lord Guest).

[16] ibid 248.

[17] ibid 249 (emphasis added). See also 249: "[A] contingent liability is … a liability which, by reason of something done by the person bound, will necessarily arise or come into being if one or more of certain events occur or do not occur." (emphasis added)

[18] ibid 254 (emphasis added). See also 253.

[19] ibid 264 (emphasis added).

[20] ibid 252.

[21] ibid 257 (emphasis added).

was not the real point of departure from the views of the majority. The actual difference of opinion turned on the nature of the triggering event, which in *Winter* was "ships being sold at a price in excess of their written down value".[22] The minority took issue with this event being in the debtor company's control, because the debtor company could have simply chosen not to sell the ships.[23] Lord Hodson held that "a state of affairs which can be terminated of one's own choice without outside intervention… [is] inconsistent with the imposition of a liability".[24] Lord Hodson found that because "the liability could have been avoided at the volition of the company… the appellants… established vulnerability… but not contingent liability."[25] Given that this was only the minority view, *Winter* stands for the proposition that a contingent liability may exist even if the debtor can choose whether to crystallise the claim or not.

B. Discretionary contingencies

Can there be a contingent debt if crystallisation is a matter of someone's discretion? In England, *Winter* demonstrates that contingent liabilities may depend on the discretion of the debtor. *Haine v Day*[26] extended this further by establishing that the discretion may be that of an unrelated party, such as a court. On the facts, what gave the protective award its contingent quality was precisely "[t]he fact that the Employment Tribunal had a discretion whether or not to make the award".[27] In the UK Supreme Court case *Re Nortel GmbH (in administration)*,[28] Lord Neuberger (with whom Lords Mance, Clarke and Toulson agreed) made a point of stating that *Haine* was "rightly decided".[29] Lord Neuberger reiterated that a contingency need not be one whose "occurrence… [is] determined by absolute rather than discretionary factors".[30] A contingent claim may even be "payable in the event that it is called upon by the creditor (in whose power it is to determine whether or not the contingency occurs)".[31] *Engwirda* demonstrates that a contingent claim in Australia may also depend upon the discretion of an unrelated party—in that case, an arbitrator.[32] Another Australian example is *McLellan v Australian*

[22] ibid 264.
[23] See also *Re T&N Ltd* [2005] EWHC 2870 (Ch), [2006] 3 All ER 697 [50] (David Richards J).
[24] *Winter* (n 13) 259.
[25] ibid, 259–260. See also *Re T&N Ltd* (n 23) [54] (David Richards J).
[26] *Haine v Day* [2008] EWCA Civ 626, [2008] 2 BCLC 517.
[27] ibid [29]. But see [55](ii), [64]–[69].
[28] *Re Nortel GmbH (in administration)* [2013] UKSC 52, [2014] AC 209.
[29] ibid [90].
[30] ibid [136].
[31] *Donnelly v Royal Bank of Scotland plc* 2016 SLT (Sh Ct) 307 [97].
[32] *Engwirda* (n 6) 458–459 (Kitto J), 461–462 (Owen J).

Stock Exchange Ltd,[33] which concerned charges laid on a company prior to it going into administration, and a fine that was imposed by the Adjudicatory Tribunal afterwards. Finkelstein J accepted that "the possibility of the imposition of a fine by a domestic tribunal is a relevant contingency"[34] and therefore held that Australian Stock Exchange Ltd was a contingent creditor.[35]

C. Differences between England and Australia

(i) How an Obligation is Assumed

Does it matter how the debtor comes under the legal obligation? The English case of *Glenister v Rowe* was concerned with whether "a person against whom a costs order may be made… [has] before an order is actually made, a 'contingent liability' for such costs".[36] Mummery LJ (with whom Thorpe and Butler-Sloss LJJ agreed)[37] held "that the claim for costs… was not a contingent liability… at the date of… bankruptcy"[38] on the basis that "it is necessary to identify something agreed or some act done by… [the potential debtor] to give rise to a liability on his part".[39] Merely being sued does not satisfy this condition, according to *Glenister*. In *Nortel*, however, Lord Neuberger expressly stated that *Glenister* was "wrongly decided".[40] He held that "[a]n order for costs made against a company in liquidation… in proceedings begun before it went into liquidation, is… provable as a contingent liability… as the liability for those costs will have arisen by reason of the obligation which the company incurred when it became party to the proceedings".[41] The judgment of Lord Sumption (with whom Lords Mance and Clarke agreed) contains similar comments.[42] Thus a contingent claim in England need not have its genesis in some independent and voluntary act of the debtor.

In the Australian High Court case of *Foots v Southern Cross Mine Management Pty Ltd*, however, Gleeson CJ, Gummow, Hayne and Crennan JJ held that a costs order made after bankruptcy was not a provable debt because there was "no antecedent

[33] *McLellan* (n 11).
[34] ibid [12]. See also [16].
[35] ibid [18]. See also *ACCC v Phoenix Institute of Australia Pty Ltd (Subject to Deed of Company Arrangement)* [2016] FCA 1246 [68]–[71] (Perry J).
[36] *Glenister* (n 14) 78–79.
[37] ibid 85.
[38] ibid 84.
[39] ibid.
[40] *Nortel* (n 28) [91].
[41] ibid [89].
[42] ibid [136].

obligation to pay costs until the order was made".[43] Although their Honours noted that, in bankruptcy, "the classes of provable debts are narrower than those… [in] corporate insolvency",[44] *Foots* has been seen to apply in the winding up context as well.[45] Yet arguably being sued brings "a person… within a system governed by rules of court, which carry with them the potential for being rendered legally liable for costs"[46]—and therefore *does* create an antecedent obligation. Rather, the fundamental difference between the *Foots* (and *Glenister*) view of a contingent claim, and the *Nortel* view on the other hand, seems to relate to *how* the antecedent obligation is assumed: in England, it may be imposed upon a potential debtor, whereas *Foots* suggests that in Australia its source must be the potential debtor's independent conduct.[47] An illustration of this is *BE Australia WD Pty Ltd (subject to a Deed of Company Arrangement) v Sutton*, where the facts involved proceedings commenced by Mrs Sutton which could have resulted in an order being made against a company, though not based on "any obligation that the company owed to her before".[48] Campbell JA (with whom McColl and Young JJA agreed) held that there was no contingent claim.[49] His Honour relied on *Foots* by drawing an analogy with costs orders, which do not depend on "any legal right that the applicant has (beyond the bare right to seek the order) before the order is actually made".[50] Yet, now that *Glenister* has been overruled by *Nortel*,[51] the claim in *Sutton* would probably qualify as a contingent claim in England. It follows that the contingent claim is a narrower concept in Australia than in England.

(ii) Mere Expectations

Contingent claims must not be confused with mere expectations. In the Australian case of *Lam Soon Australia Pty Ltd (administrator appointed) v Molit (No 55) Pty Ltd*, the Full Federal Court commented in obiter that "[a] right to sue for damages for a particular future breach of… covenant,… looked at *before the breach occurs*, [is] not

[43] *Foots v Southern Cross Mine Management Pty Ltd* (2007) 234 CLR 52 [65]. See also [35]–[36], quoting and citing *Glenister* (n 14).

[44] ibid [9].

[45] See, for example, *Central Queensland Development Corp Pty Ltd v Sunstruct Pty Ltd* (2015) 231 FCR 17 [42]–[72] (Gilmour J, Rangiah and Besanko JJ agreeing); Sutton (n 11) [114] (Campbell JA, McColl and Young JJA agreeing); *Larkden Pty Ltd v Lloyd Energy Systems Pty Ltd* [2011] NSWSC 1567 [61], [69] (Hammerschlag J). But see *Re Walker* (2007) 215 FLR 428 [18] (Barrett J).

[46] *Nortel* (n 28) [89].

[47] See also *McCluskey v Pasminco Ltd* (2002) 120 FCR 326, [39]–[41] (Goldberg J).

[48] *Sutton* (n 11) [116]. See also [31].

[49] ibid [144].

[50] ibid [115]. See generally [113]–[117].

[51] *BPE Solicitors v Gabriel* [2015] UKSC 39, [2015] AC 166 [13], citing Nortel (n 28) [87]–[93] (Lord Neuberger), [136] (Lord Sumption).

even a contingent claim: it is a mere expectancy and could not be the subject of proof".[52] Although contractual obligations are voluntarily assumed, perhaps this statement was motived by a concern that the potential debtor has control over whether a breach will or will not occur (reminiscent of the minority's concern in *Winter*). Conversely, as Lord Hoffmann noted in *Secretary of State for Trade and Industry v Frid*, an established category of contingent debts in England comprises of "claims for breach of contract where the contract was made before the insolvency date but the breach occurred afterwards".[53] In an Australian case of the same year, *Re National Express Group Australia (Swanston Trams) Pty Ltd*, Finkelstein J staunchly disagreed with *Lam Soon* and relied on English authorities to argue that "a future breach of contract could be proved as a contingent claim".[54] Finkelstein J's remarks were, however, obiter.[55] The Australian position remains unclear.[56] Thus, the comment in *Lam Soon* creates another possible basis on which contingent claims may be a narrower concept in Australia than in England.

D. Examples

The differences between a contingent claim in England and in Australia mean that the scope of the problem identified in Part III may be slightly broader in the former jurisdiction. To the extent that the problem creates unfairness—explored in Part IV below—one would expect more instances of unfairness in England than in Australia, *ceteris paribus*. Yet it is also apparent from the above discussion that the concepts of contingent claims in these jurisdictions largely overlap. Contracts are a common source of contingent claims.[57] Putting potential breaches aside, one typical example is "an uncalled guarantee, [because] the person with the benefit of the guarantee will be a 'creditor' of the guarantor, even though there has not been, and may never be, any default on the principal debt or any call on the guarantee".[58]

[52] *Lam Soon Australia Pty Ltd (administrator appointed) v Molit* (No 55) Pty Ltd (1996) 70 FCR 34, 44 (von Doussa, O'Loughlin and Lehane JJ) (emphasis added).

[53] *Secretary of State for Trade and Industry v Frid* [2004] UKHL 24, [2004] 2 AC 506 [9], citing *Re Asphaltic Wood Pavement Co* (1885) 30 Ch D 216 (CA).

[54] *Re National Express Group Australia (Swanston Trams) Pty Ltd; Thiess Infraco (Swanston) Pty Ltd v Smith* (2004) 209 ALR 694 [10]. See generally [10]–[16].

[55] ibid [17].

[56] See *Australian Gypsum* (n 11) [75], [79] (Buss JA); *Larkden* (n 45) [58] (Hammerschlag J); *Re Walker* (n 45) [19] (Barrett J); *Spring Hill Apartments Pty Ltd v We Both Pty Ltd* (in liq) [2006] QSC 151 [40] (Moynihan J); *Re Motor Group Australia Pty Ltd* (2005) 54 ACSR 389 [8]–[11], [14] (Hely J); *Wallace-Smith v Thiess Infraco* (Swanston) Pty Ltd (2005) 218 ALR 1 [105]–[106] (French J), [341] (Allsop J).

[57] See, for example, *Nortel* (n 28) [75] (Lord Neuberger), [131] (Lord Sumption).

[58] *Re T&N Ltd* (n 23) [46] (David Richards J).

Other common examples include possible claims under insurance policies[59] and conditional fee agreements.[60] Importantly, claims—including contingent claims—"need not arise from consensual transactions at all, but can arise from statutory obligations… or torts".[61] In the English case of *Re T&N Ltd*, David Richards J had to decide whether "those persons who have been exposed to asbestos and who will have claims in negligence… *if they develop* asbestos-related diseases, are 'creditors'… for the purposes of [restructuring regimes]".[62] The answer was in the affirmative, at least in circumstances where "the relevant acts or omissions… are complete… [and] the potential claimants have been exposed to asbestos".[63]

III. The Jekyll and Hyde Problem

The easiest way to understand the problem[64] is to consider some simple examples. If a £100 debt is due and payable when a winding up commences in England, the creditor would be entitled to prove for £100 plus any accrued interest.[65] (In Australia, a creditor who is owed $100 could likewise prove for $100 plus any interest.[66]) Next, consider a future debt of £100 that will fall due three years from now. If interest rates are positive, £100 today is worth more than the same amount in the future, so the debt must be discounted because of "the time value of money".[67] Assuming that the anticipated average interest rate over the forthcoming three years is 5% per annum, the present value of the debt is £86.38.[68] But what if the actual average rate of interest turns out to be only 1% per annum over those three years? Then the discount rate was too high—to the detriment of the future creditor, who *ought* to have been permitted to prove for £97.06.[69] Conversely, if future interest rates are underestimated, discount rates are too low, and future claims are overvalued. This unduly benefits our hypothetical future creditor, who can invest the sum paid through the liquidation or restructuring

59 *Winter* (n 13) 258 (Lord Hodson).
60 *Rowbury v Official Receiver* [2015] EWHC 2276 (Ch), [2016] BPIR 477 [26].
61 *Frid* (n 53) [25] (Lord Hoffmann). See also [19]; *Re T&N Ltd* (n 23) [63].
62 *Re T&N Ltd* (n 23) [66] (emphasis added). See also [47].
63 ibid [67]. See also *Secretary of State for Business, Innovation and Skills v Broomfield Developments Ltd* [2014] EWHC 3925 (Ch) [49]–[51], [80]–[81].
64 Part III expands upon the author's previous work concerning this problem: see n 5.
65 Insolvency (England and Wales) Rules 2016, r 14.23(1)–(6).
66 Corporations Act 2001 (Cth), s 554(1).
67 *Lomas v Burlington Loan Management Ltd* [2015] EWHC 2269 (Ch), [2016] Bus LR 17 [215] (David Richards J) (*Lomas*).
68 Present value = (future amount)/(1 + interest rate)years = £100/(1 + 0.05)3 ≈ £86.38.
69 Present value = (future amount)/(1 + interest rate)years = £100/(1 + 0.01)3 ≈ £97.06.

process, and in three years' time can have more than the £100 that was owed to the creditor by the company.

For contingent claims, one must first assess "the chances of the contingency occurring and the likely amount of any claim"[70] and then assign "a present value on possible future events or outcomes."[71] Once again, "if the contingent debt cannot fall due for payment for a period of, say, five years, the estimate of the liability must include an element of *discount for that period*".[72] To take a very simple example, suppose there is a 50% chance of a £100 claim crystallising three years from now, and the average interest rate over that period is expected to be 5% per annum. The present value of this contingent claim is £43.19.[73] Yet consider the possible outcomes in this scenario if the creditor is involved in the debtor company's liquidation or restructuring scheme or arrangement:

TABLE III.1

Claim will	*Contingent creditor receives up to £43.19 today*
Arise	Yet would have been owed £100 in three years, the present value of which is £86.38
Not arise	Yet would not have been owed anything at all

Thus, a contingent claim which *will* subsequently crystallise is undervalued at present.[74] But if it will *not* subsequently crystallise, the debtor company is paying someone who will never be owed anything at all![75]

Clearly there is a difficulty with assigning a present value to claims that are not due and payable as yet. Specifically, future and contingent claims create a problem in winding up and restructuring processes because their value must be *estimated*. In the case of future claims, the discount rate for the time value of money has been legislatively fixed at 5% per annum in England,[76] and 8% per annum in Australia.[77] Although there is no risk of paying someone whose claim will never crystallise, these rates may in time prove to have been too high or too low, meaning that future claims are undervalued or overvalued today. There is an additional

[70] *Re MF Global UK Ltd (in special administration) (No. 2)* [2013] EWHC 92 (Ch), [2013] Bus LR 1030 [76] (David Richards J).

[71] ibid [54] (David Richards J).

[72] *Lomas* (n 67) [198] (David Richards J) (emphasis added).

[73] Present value = (future amount x probability)/(1 + interest rate)years = (£100 x 0.50)/(1 + 0.05)3 ≈ £43.19.

[74] See Thomas H Jackson, *The Logic and Limits of Bankruptcy Law* (reprint, Beard Books 2001) 47.

[75] See *Lomas* (n 67) [203] (David Richards J); *Wight v Eckhardt Marine GmbH* [2003] UKPC 37, [2004] 1 AC 147 [29], [33] (Lord Hoffman)

[76] Insolvency (England and Wales) Rules 2016, r 14.44(2).

[77] Corporations Regulations 2001 (Cth), reg 5.6.44. See also Corporations Act 2001 (Cth), s 554B.

estimation problem in the case of contingent claims, given that they must also be discounted for the probability of non-crystallisation. Like Dr Jekyll, who concealed his transformations from the world, future and contingent claims may be overestimated; or, like Mr Hyde, who was but one persona adopted by a more complex character, future and contingent claims may be underestimated.[78] Only time can tell which of the two possibilities will become reality. This valuation—or 'Jekyll and Hyde'—problem is the consequence of involving future and contingent creditors in liquidation and restructuring processes.

IV. Unfair Distributions and Votes

This Part of the article investigates two ways in which the Jekyll and Hyde problem may cause unfairness.[79] For one, it may produce unfair distributions, not only to the future or contingent creditors themselves, but also to its shareholders or to the company's creditors *as a whole*. Secondly, it may produce unfair voting if creditors are invited to make a decision about the future of the debtor company. In contending that distributions and voting may be unfair when future and contingent creditors are involved, the article draws on the reasoning that underpins an established principle in English and Australian insolvency law: namely, the 'hindsight principle'.

A. Distributions

(i) Skewed Distributions

How are the assets of a company distributed in a winding up? In the voluntary winding up of an English company, its assets are generally "applied in satisfaction of the company's liabilities *pari passu* and… [then] distributed among the members according to their rights and interests in the company".[80] Similarly, solvent proprietary companies in Australia limited by shares must pay all their debts in full and any surplus must be distributed to shareholders.[81] But if there are not enough assets to meet all the claims of creditors, the *pari passu* principle requires "the free assets of the insolvent… [to] be distributed *rateably* amongst the insolvent's unsecured creditors"[82]—in other words, "*pro rata* according to the

[78] Robert Louis Stevenson, The Strange Case of Dr Jekyll and Mr Hyde (Longmans, Green & Co 1886).

[79] Part IV expands upon the author's previous work concerning these two forms of unfairness: see n 5.

[80] Insolvency Act 1986 (UK), s 107.

[81] See, for example, Corporations Act 2001 (Cth), Pt 1.5, para 9.4; *Mills v Sheahan* (2007) 99 SASR 357 [13] (Debelle J); *Commissioner of Taxation v Linter Textiles Australia Ltd* (in liq) (2005) 220 CLR 592 [54] (Gleeson CJ, Gummow, Hayne, Callinan and Heydon JJ), [121] (McHugh J).

[82] *Re Gray's Inn Construction Co Ltd* [1980] 1 All ER 814 (CA) 819 (Buckley LJ) (emphasis added).

value of… [their] claims".[83] In both jurisdictions, this "basic concept of [the] law governing the liquidation of insolvent estates"[84] is enshrined in statute. For instance, the rule applies in an English administration or a court winding up in the same way as it does in a voluntary winding up:

> Debts other than preferential debts rank equally between themselves and, after the preferential debts, must be paid in full unless the assets are insufficient for meeting them, in which case they abate in *equal proportions* between themselves.[85]

Similarly in Australia, subject to certain exceptions, "all debts and claims proved in a winding up rank equally and, if the property of the company is insufficient to meet them in full, they must be paid *proportionately*".[86]

When "the available pot is too small to pay everyone in full, a *pari passu* distribution has an obvious appeal".[87] It has the virtue of simplicity.[88] More importantly, "*pari passu* distribution… responds to a very basic human feeling that, when faced by a common misfortune, all those affected by it should bear the burden equally".[89] Indeed, "*[p]ari passu* distribution is derived from the maxim that 'equality is equity'".[90] Thus Briggs J in *Re Nortel GmbH (in administration)* called it "a

[83] *Re Dynamics Corporation of America* [1976] 2 All ER 669 (Ch) 673 (Oliver J) (emphasis added). See also *Fuglers LLP v Solicitors Regulatory Authority* [2014] EWHC 179 (Admin), [2014] BPIR 610 [18] (Popplewell J); *Revenue and Customs Commissioners v Football League Ltd* [2012] EWHC 1372 (Ch), [2012] Bus LR 1539 [4] (David Richards J).

[84] *Re Gray's Inn Construction Co Ltd* (n 82) 819. See also, for example, *Coshott v Barry* (2015) 91 NSWLR 1 [88] (McColl and Emmett JJA and Brereton J); *Westpac Banking Corporation v Bell Group Ltd* (in liq) (No 3) (2012) 44 WAR 1 [792] (Lee AJA); *Re Opes Prime Stockbroking Ltd* (2008) 171 FCR 473 [7] (Finkelstein J) ; *Edwards v Attorney General* (2004) 60 NSWLR 667 [89] (Young CJ in Eq); Lewis v Hyde [1998] 1 WLR 94 (PC), 98 (Lord Browne-Wilkinson); *Re HIH Casualty and General Insurance Ltd* (n 3) [34], [37] (David Richards J); *Re HIH Casualty and General Insurance Ltd* (HL) (n 3) [2] (Lord Hoffman).

[85] Insolvency (England and Wales) Rules 2016, r 14.12(2) (emphasis added). See also Insolvency Act 1986 (UK), ss 107 (voluntary winding up), 175(1A)–(1B) (preferential debts in winding up), 328 (bankruptcy).

[86] Corporations Act 2001 (Cth), s 555 (emphasis added). See also Bankruptcy Act 1966 (Cth), ss 108 and 109(11).

[87] *Re Sigma Finance Corporation (in Administrative Receivership)* [2008] EWCA Civ 1303 [92] (Rimer LJ).

[88] *Barlow Clowes International Ltd (in liq) v Vaughan* [1992] 4 All ER 22 (CA) 36 (Woolf LJ).

[89] *Charity Commission for England and Wales v Framjee* [2014] EWHC 2507 (Ch), [2015] 1 WLR 16 [61] (Henderson J). See also, for example, *Re Lehman Brothers International (Europe) (in admin)* [2010] EWCA Civ 917, [2011] Bus LR 277 [76] (Arden LJ); *Re Golden Key Ltd (in receivership)* [2009] EWCA Civ 636 [63] (Arden LJ).

[90] *Re Golden Key Ltd (in receivership)* (n 89) [3] (Arden LJ). See also, for example, *Barlow Clowes* (n 88) 42 (Woolf LJ).

fundamental principle of justice, equity and fairness".[91] In Australia, the Full Federal Court in *Akers v Deputy Commissioner of Taxation* recognised that "[t]he principle of *pari passu* distribution… is informed by fairness and equality".[92] The *pari passu* rule, however, does not deviate from the requirement that all provable debts be valued at the same point in time. In *Re Dynamics Corporation of America*, Oliver J explained that "the claims of the creditors amongst whom the division is to be effected must all be *crystallised at the same date*, even though the actual ascertainment may not be possible at that date, for otherwise one is not comparing like with like".[93] The Privy Council in *Wight v Eckhardt Marine GmbH* reiterated that "valuation at the date of winding up ensures that distribution among creditors is truly *pari passu*."[94]

Returning to the Jekyll and Hyde problem outlined in Part III, the overvaluation of future and contingent debts will result in overpayment to future and contingent creditors in a winding up. All else being equal, this overpayment will be detrimental to the members of a solvent company because the surplus available to be distributed among them will be reduced. Or if the company is insolvent, it is the other creditors who will be disadvantaged. This is because (*ceteris paribus*) as the demands on an insufficient pool of assets increase, the portion of each creditor's claim that can be satisfied by the company's limited assets is reduced.[95] Conversely, underpayment to future and contingent creditors unduly benefits members (of a solvent company) or the other creditors (of an insolvent company). This was apparent by 1888, when the Earl of Selborne (with whom Lords Fitzgerald and Herschell agreed) astutely commented in *Hardy v Fothergill*:

> For the principle of the old bankrupt laws, which did not admit to proof any claims for liabilities contingent at the time of bankruptcy, much might perhaps be said. It might be a *hardship upon the creditors*, to whom debts were then due, that a merely possible future liability, *which might never have matured into a debt at all*, should be valued and admitted to proof and to *participation in dividends*.… [B]ut there might also be cases in which it would be a *hardship upon him to be compelled to*

[91] *Re Nortel GmbH (in administration)* [2010] EWHC 3010 (Ch), [2011] Bus LR 766 [64]. See also, for example, *Re Lehman Brothers International (Europe) (in admin)* [2009] EWHC 3228 (Ch), [2010] 2 BCLC 301 [249] (Briggs J); Wight (n 75) [28]; *Russell-Cooke Trust Co v Prentis* [2002] EWHC 2227 (Ch), [2003] 2 All ER 478 [58] (Lindsay J); *Re Polly Peck International plc (in administration) (No 2)* [1998] 3 All ER 812 (CA) 826 (Mummery LJ); *Cox v Bankside Members Agency Ltd* [1995] CLC 671, 682 (Peter Gibson LJ).

[92] *Akers v Deputy Commissioner of Taxation* (2014) 223 FCR 8 [138] (Allsop CJ, Robertson and Griffiths JJ agreeing). See also, for example, *Re Opes Prime Stockbroking Ltd* (n 84) [7] (Finkelstein J); *Sheahan v Carrier Air Conditioning Pty Ltd & Campbell* (1997) 189 CLR 407, 463–464 (Kirby J).

[93] Re Dynamics Corporation of America (n 83) 675–676 (emphasis added).

[94] *Wight* (n 75) [29].

[95] See, for example, *Re Trident Fashions plc* [2004] EWHC 293 (Ch), [2004] 2 BCLC 35 [30] (Lewison J); *Re FMS Financial Management Services Ltd* (1989) 5 BCC 191, 193 (Hoffmann J).

> *prove such a claim, so valued*, at a time when the estate of the bankrupt might be inadequate to pay any substantial dividend, instead of waiting for the occurrence of the contingency (if it ever should occur) and then taking his legal remedies against such property as the debtor might at that time possess, whether previously bankrupt or not.[96]

(ii) The Hindsight Principle

The Jekyll and Hyde problem plainly produces skewed distributions—not only for future and contingent creditors, but also for members or other creditors—in English and Australian liquidations. This is particularly noticeable when future and contingent claims mature after the commencement of a winding up, but *prior* to any payment being made. Returning to the example of a £100 future debt, even if it becomes due and payable by the time of distribution, the requirement to value all claims as at a single date would mean that our hypothetical creditor could rely only on a proof for £86.38.[97] Yet once a future debt matures, it seems "unjust that this creditor should not receive a dividend on the full amount of his debt".[98] Again, suppose that a contingent claim crystallises prior to any distribution: if £100 is due and payable, why should the creditor's proof be limited to something like £43.19?[99] Or if it becomes certain that the contingent claim will never crystallise, why should the company pay anything at all?[100] The injustice is so glaring that courts have developed an exception to the same-date valuation requirement: the hindsight principle.[101]

The hindsight principle permits claims to be valued at what they *are* or *ought to have been* worth, rather than what they *were in fact* worth at the commencement of proceedings.[102] As David Richards J explained in *Re MF Global UK Ltd (in special administration)*, it "either removes the need to make the estimate or makes the estimate more accurate and produces what may generally be regarded as *fairer*

[96] *Hardy v Fothergill* (1888) 13 App Cas 351, 358 (emphasis added).

[97] See n 68 and accompanying text.

[98] *Lomas* (n 67) [215] (David Richards J).

[99] See n 73 and accompanying text.

[100] See n 75.

[101] See, for example, *Grapecorp Management Pty Ltd (in liq) v Grape Exchange Management Euston Pty Ltd* (2012) 265 FLR 33 [68]–[70] (Sifris J); *New Cap Reinsurance Ltd (in liq) v Grant* (2008) 221 FLR 164 [50]–[52] (White J); *MS Fashions Ltd v Bank of Credit and Commerce International SA* (in liq) (No 2) [1993] Ch 425 (CA) 432, 435 (Hoffmann LJ).

[102] See *Lomas* (n 67) [200], [205] (David Richards J); *Gleave v Board of the Pension Protection Fund* [2008] EWHC 1099 (Ch), [2008] Bus LR 1443 [20] (David Richards J); *Stein v Blake* [1996] AC 243, 252 (Lord Hoffmann).

values for the purposes of the distribution or payment".[103] The Privy Council similarly held in *Wight*:

> Hindsight is used because it is *not considered fair* to a creditor to value a contingent debt at what it might have been worth at the date of the winding-up order when one knows that prescience would have shown it to be *worth more*. The same must be true of a contingent debt which prescience would have shown to be *worth less*.[104]

The *raison d'être* of the hindsight principle is the idea that overvaluing and undervaluing uncrystallised claims is unfair because it results in distributions to future and contingent creditors that are either too large, or too small—and this unduly harms or benefits other creditors of the insolvent company, or members of the solvent company.

The hindsight principle is now "expressly recognised in the applicable legislation".[105] In England, a future claim must be discounted only if "payment is not due at the date of the declaration of a dividend".[106] Likewise in Australia, "[t]he discount by which the amount payable on the future date is to be reduced… is… calculated from the declaration of the dividend to the time when the debt would have become payable".[107] Thus future claims are not discounted if they fall due before a dividend is declared.[108] Yet the hindsight principle is temporally limited: the Jekyll and Hyde problem remains in respect of future claims which do not mature before a dividend is declared. Regarding contingent claims, in an English "administration or… winding up, the office-holder must estimate the value of a debt that does not have a certain value".[109] Consistently with the hindsight principle, such an estimate may be revised "by reference to a change of circumstances or to information becoming available to the office-holder."[110] Thus "the amount provable… is that of the estimate *for the time being*".[111] Additionally, a creditor may withdraw a proof at any time, or vary the amount by agreement with the office-holder.[112] Similarly, in an Australian winding up, "[a] proof of debt or claim may be withdrawn, reduced or varied by a creditor with the consent

[103] *Re MF Global UK Ltd (in special administration)* [2013] EWHC 92 (Ch), [2013] Bus LR 1030 [48] (emphasis added).
[104] *Wight* (n 75) [32] (emphasis added). See also *Stein* (n 103) 252 (Lord Hoffmann).
[105] *Re MF Global UK Ltd (in special administration)* (n 104) [52] (David Richards J).
[106] Insolvency (England and Wales) Rules 2016, r 14.44(1).
[107] Corporations Regulations 2001 (Cth), reg 5.6.44. See also Corporations Act 2001 (Cth), s 554B.
[108] See also *Lomas* (n 67) [197], [224], regarding repealed Insolvency Rules 1986 (UK), r 2.105.
[109] Insolvency (England and Wales) Rules 2016, r 14.14(1).
[110] ibid r 14.14(2).
[111] ibid r 14.14(4) (emphasis added).
[112] ibid r 14.10.

of the liquidator".[113] Such provisions enable a contingent creditor to "come back for more if the contingency eventuated".[114] There is no possibility, however, of variation or revision after liquidation,[115] given that "previous distributions cannot be set aside".[116] For contingent claims that do not become certain in time, the Jekyll and Hyde problem remains unresolved.

B. Voting on alternatives

Two corporate restructuring mechanisms that exist in England and Australia in similar forms—voluntary arrangements and court-sanctioned schemes—are considered below. These regimes typically involve creditors voting on particular proposals for dealing with the company's debts. After outlining how the proposals may be approved, this section will elucidate the way in which the Jekyll and Hyde problem infects the voting processes for both of these restructuring regimes.

(i) Voluntary Arrangements

In England, "[t]he directors of a company… may make a proposal… to the company and to its creditors for a composition in satisfaction of its debts or a scheme of arrangement of its affairs".[117] A decision must be sought "from the company's creditors as to whether they approve the proposal".[118] Accordingly, an invitation will be sent to the company's creditors to participate in some kind of a qualifying decision procedure.[119] A decision approving a proposed or modified company voluntary arrangement "is made when three-quarters or more *(in value)* of those responding vote in favour of it".[120] "Votes are calculated according to the amount of each creditor's claim"[121] at the relevant date, with two major exceptions. First, the value of a secured creditor's vote is usually limited to the value of any unsecured part; and if there is no unsecured part, it is nil.[122] Secondly, "[a] creditor may vote in respect of a debt of an *unliquidated* or *unascertained* amount if the

[113] Corporations Regulations 2001 (Cth), reg 5.6.56. See also Corporations Act 2001 (Cth), s 554A.
[114] Re *Lehman Brothers International (Europe) (in administration) (No 4)* [2015] EWCA Civ 485, [2016] Ch 50 (CA) [95] (Lewison LJ), citing *Re Danka Business Systems plc (in members' voluntary liquidation)* [2013] EWCA Civ 92, [2013] Ch 506 (CA) [37] (Patten LJ).
[115] See *Danka* (n 114) [21], [37].
[116] *Wight* (n 75) [31] (Lord Hoffman).
[117] Insolvency Act 1986 (UK), s 1(1). But see also s 1(3).
[118] ibid ss 3(1)(b) and 3(2)(b). See also Sch A1, para 29(1)(b). See further s 4(1A).
[119] ibid s 3(3); Sch A1, para 29(2); Insolvency (England and Wales) Rules 2016, r 2.25(2).
[120] Insolvency (England and Wales) Rules 2016, r 15.34(3)(a) (emphasis added). See also rr 15.34(4) and 15.34(5)(a).
[121] ibid r 15.31(1)(d).
[122] ibid r 15.31(4)–(6).

convener or chair decides to put upon it an *estimated minimum value* for the purpose of entitlement to vote and admits the claim for that purpose".[123] In relation to voluntary arrangements, however, "a debt of an unliquidated or unascertained amount is to be *valued at £1* for the purposes of voting unless the convener or chair or an appointed person *decides to put a higher value* on it".[124]

In Australia, the board of directors may resolve that an administrator should be appointed if it also resolves that "the company is insolvent, or is likely to become insolvent at some future time".[125] Meetings must be convened by the administrator,[126] including one at which the company's creditors can decide to end the administration, to wind up the company, or that the company execute a deed of company arrangement.[127] A resolution put to the vote of creditors must be decided on the voices, unless a poll is demanded.[128] If a poll is taken at a creditors' meeting, the resolution will be carried if a numerical majority of the creditors voting are in favour of it, and "the *value* of the debts owed by the corporation to those voting in favour of the resolution is more than half the total debts owed to all the creditors voting".[129] (But only one of these matters will suffice if "the person presiding at the meeting… exercise[s] a casting vote in favour of the resolution".[130]) A secured creditor's vote is not[131] limited to the "balance, if any, due to him or her after deducting the value of his or her security".[132] However, a "creditor must not vote in respect of… an unliquidated… or… a contingent debt… or an unliquidated or a contingent claim… or… a debt the value of which is not established… unless a just estimate of its value has been made".[133] If there is a creditor with such a debt or claim, "there is an obligation on the chairman to make… a just estimate".[134] Once this has been done, "the admission of the proof is necessarily for the amount of that just estimate".[135]

(ii) Schemes of Arrangement

There is another restructuring mechanism in England; one which involves creditors, or members, or both, voting on a proposed compromise or arrangement

[123] ibid r 15.31(2) (emphasis added).
[124] ibid r 15.31(3) (emphasis added).
[125] Corporations Act 2001 (Cth), s 436A(1). But see also ss 436A(2), 436B, 436C(1)–(2).
[126] ibid ss 436E and 439A.
[127] ibid s 439C.
[128] Corporations Regulations 2001 (Cth), reg 5.6.19(1).
[129] ibid reg 5.6.21(2) (emphasis added). See also regs 5.6.21(4)(a) and 5.6.21(4B).
[130] ibid reg 5.6.21(4)(a).
[131] ibid reg 5.6.24(4).
[132] ibid reg 5.6.24(2).
[133] ibid reg 5.6.23(2).
[134] *Kirwan v Cresvale Far East Ltd* (in liq) (2002) 44 ACSR 21 [395] (Young CJ in Eq).
[135] *Re Free Wesleyan Church of Tonga in Australia Inc* (2012) 260 FLR 348 [37] (Black J). See also [16].

in classes.[136] Upon application, an English court may sanction a compromise or arrangement "[i]f a majority in number representing 75% *in value of the creditors or class of creditors*… (as the case may be), present and voting either in person or by proxy… agree".[137] A similar mechanism is available in Australia.[138] For such a scheme between a company and its creditors (or a class of creditors) to be approved there must be a favourable vote from a numerical majority and their "debts or claims against the company [must] amount in the aggregate to at least 75% of the *total amount of the debts and claims of the creditors present and voting* in person or by proxy, or of the creditors included in that class present and voting in person or by proxy, as the case may be".[139]

(iii) Value of Vote

Earlier, this Part demonstrated that the Jekyll and Hyde problem produces skewed distributions when future and contingent creditors are involved in a winding up. Importantly, it also produces distortions whenever these creditors are involved in voting on the debtor company's future: the value of a creditor's vote in respect of a proposed voluntary arrangement or scheme normally corresponds to the value of that creditor's debt or claim.[140] Therefore, future and contingent creditors will not only receive too much or not enough from the debtor company, to the detriment or benefit of others, but antecedent to any distribution, their votes under a restructuring regime will either be too weighty or too light. For future creditors, the voting unfairness is proportionate to the difference between the actual and estimated interest rate. In the case of contingent creditors, if their claims will eventually crystallise, they deserve more voting power; and if their claims will never crystallise, they should not be voting at all. For contingent creditors whose claims *will* crystallise, the unfairness of insufficient voting power will be greatly exacerbated under the English voluntary arrangement regime if their claim is valued at £1 for the purposes of voting on a proposal.[141] Similarly in Australia, a creditor's voting entitlement might legitimately be reduced to as little as $1 if the "debt is subject to an uncertain contingency".[142]

[136] Companies Act 2006 (UK), Part 26, especially ss 895(1)(a) and 896(1).
[137] ibid s 899(1) (emphasis added).
[138] Corporations Act 2001 (Cth), Part 5.1, especially s 411.
[139] ibid s 411(4)(a)(i) (emphasis added).
[140] See nn 120, 121, 129, 137, 139 and accompanying text.
[141] See n 124 and accompanying text. See also *Re Gertner* [2017] EWHC 111 (Ch), [2017] BPIR 336 [73] (Judge Keyser QC); *Revenue and Customs Commissioners v Maxwell* [2010] EWCA Civ 1379, [2011] Bus LR 707 [57] (Lord Neuberger MR, Carnwath and Sullivan LJJ agreeing).
[142] *Selim v McGrath* (2003) 177 FLR 85 [269](f) (citations omitted) (Barrett J). See also nn 133–135 and accompanying text.

V. Conclusion

Future and contingent claims are valued differently from claims that are already due and payable because they are discounted. Future claims are discounted to take into account the time-value of money—which gives rise to overestimation or underestimation of their present value, depending on whether the discount rate proves too low or too high. Contingent claims are further discounted for the uncertainty of crystallisation. This results in an undervaluation of contingent claims which will crystallise, or in the treatment of people and entities as 'creditors' though they will never become entitled to payment at all. This difficulty with ascertaining the value of uncrystallised debts and claims was dubbed the Jekyll and Hyde problem in Part III. That it is unfair is already implicitly recognised in insolvency law by the hindsight principle, though it is temporally limited. The problem manifests itself in how much future and contingent creditors (as well as members or other creditors, depending on the debtor company's solvency) receive in a winding up. Additionally, the problem may manifest itself when creditors vote on restructuring proposals. Indeed, the distributions offered to creditors under such proposals are likely to be indirectly distorted by the Jekyll and Hyde problem, to the extent that winding up is the appropriate counterfactual to the restructure.[143] Yet plainly future and contingent creditors must continue to participate in liquidations, given that companies' affairs are being finalised.[144]

Nonetheless, the Jekyll and Hyde problem is not entirely intractable. In relation to future claims, the discount rate is presently fixed by the legislatures in England and Australia (at 5% and 8%, respectively).[145] It is respectfully submitted that the problem can be largely resolved by simply replacing these fixed rates with floating ones that are pegged to some relevant rate in each jurisdiction's marketplace, such as an appropriate fixed-term deposit rate. Variable discount rates may still prove too low or too high by the time that a future claim matures, but are likely to produce more accurate valuations. In turn, better valuations will reduce the unfairness associated with future creditors voting on the future of, and

[143] See, in relation to voluntary arrangements: *Mourant & Co Trustees Ltd v Sixty UK Ltd (in liq)* [2010] EWHC 1890 (Ch), [2011] 1 BCLC 383 [67](c) (Henderson J); *Fleet Broadband Holdings Pty Ltd v Paradox Digital Pty Ltd* (2005) 228 ALR 598 [62] (Master Newnes); *Infa Products Pty Ltd v Infa Products Pty Ltd (Administrator Appointed)*; *Re Britax Childcare Pty Ltd* (2016) 115 ACSR 322 [115] (Burley J). See, in relation to court-sanctioned schemes: *Re T&N Ltd* [2004] EWHC 2361 (Ch), [2005] Pens LR 1 [82] (David Richards J); *Re Pheon Pty Ltd* (1986) 47 SASR 427, 443 (White J).

[144] See, for example, *Unite the Union v Nortel Networks UK Ltd* (in administration) [2010] EWHC 826 (Ch), [2010] 2 BCLC 674 [33] (Norris J); *Mine & Quarry Equipment International Ltd v McIntosh* (2005) 54 ACSR 1 [16] (McPherson JA, Atkinson and Mullins JJ agreeing); *Financial Services Compensation Scheme Ltd v Larnell (Insurances) Ltd (in liq)* [2005] EWCA Civ 1408, [2006] QB 808 (CA) [19] (Lloyd LJ), [58] (Moore-Bick LJ); *Re T & N Ltd* (n 23) [94] (David Richards J).

[145] Insolvency (England and Wales) Rules 2016, r 14.44(2); Corporations Regulations 2001 (Cth), reg 5.6.44; Corporations Act 2001 (Cth), s 554B.

receiving distributions from, the debtor company. Unfortunately, in relation to contingent claims, there is no presently-available objective factor that can reduce the uncertainty of crystallisation. Contingent creditors (such as potential victims of a mass tort)[146] might not, however, want to see "very small amounts being paid to a very large number of people".[147] The Jekyll and Hyde problem's impact on distributions may be ameliorated if the company were "to calculate on a statistical and actuarial basis the sum needed to meet future claims and allow a proof for that total sum, thereby creating a reserve which can be applied in paying a dividend to those [whose claims crystallise]".[148] Undoubtedly, there would be costs associated with maintaining and distributing the reserved funds. Yet this may be a fairer way to proceed in some liquidations or under some voluntary arrangements and schemes.

[146] *Re Centro Properties Ltd* (2011) 86 ACSR 584 [21] (Barrett J). See also, for example, *New Cap* (n 101) [79] (White J); *Chief Commissioner of State Revenue v Reliance Financial Services Pty Ltd* [2006] NSWSC 1017 [36] (White J); *Petrochemical Industries Ltd (in liq) v Dempster Nominees Pty Ltd* (1994) 15 ACSR 468, 475–476 (Murray J); *Re Gasbourne Pty Ltd* [1984] VR 801, 837 (Nicholson J).

[147] *Re T & N Ltd* (n 23) [138] (David Richards J).

[148] ibid.

Feeding Decisions at the End-of-Life: Law, Ethics and Emotions

Kirsty McKenzie*

I. Introduction

The act of feeding holds considerable emotional sentiment. It is symbolic in expressing love, compassion, and nurture. Feeding is essential to our existence, not just as a means of physical sustenance, but also because of our social need to nurture and feed others. Medical interventions can frustrate these needs. This was demonstrated in 2013, when the British Dietetics Association (BDA) made a policy regarding the use of home-made liquidised food for tube-fed patients. There had been a rise in requests for advice, which was partly attributed to carers seeing it as a way to "reconnect with caring".[1] The BDA advised against home-made food because it increases the likelihood of feeding tube blockages and gastric infections.[2]

Medical advances in clinically assisted nutrition and hydration have created ethical dilemmas about what its role should be at the end-of-life. Many families see clinically assisted nutrition and hydration as a means of improving their loved

* Highly Specialist Speech and Language Therapist, Lishman Unit, Bethlem Royal Hospital. B.Sc. Hons (City University), M.A. (Medical Ethics and Law) (King's College London). I am grateful to the anonymous reviewers at the Cambridge Law Review; and give special thanks to Jillian Craigie, Penney Lewis, Alice Smith, and Pat Walsh for their guidance and support.

1 The British Dietetic Association, 'Policy Statement: Use of Liquidised Food with Enteral Feeding Tubes' (2013) 2 <www.bda.uk.com/improvinghealth/healthprofessionals/policystatement_liquidisedfood> accessed 15 August 2017.

2 ibid 3.

one's comfort at the end-of-life, despite contrary medical evidence and opinion.[3] This has sometimes been permitted, as it is acknowledged that "symbolic feeding" serves the family's means of fulfilling their duty to nourish, despite having little or no benefit to the patient.[4] These observations confirm the important emotional component of feeding. This calls for further investigation of the role of emotions in legal decisions regarding feeding.

Two key areas of law where this issue arises are end-of-life decisions in the context of prolonged disorder of consciousness (PDCs) and anorexia nervosa (hereinafter, "anorexia"). This article uses an analysis of legislation and cases in these two areas of law to explore how emotion is manifested, and what influence it may have had on judicial reasoning. I will consider specific cases, but also reflect on trends since the landmark rulings of *Airedale NHS Trust v Bland* in relation to PDCs,[5] and *Re E* in relation to anorexia.[6] Anthony 'Tony' Bland was eighteen years old when he was crushed in the Hillsborough disaster, inflicting him with a prolonged disorder affecting his consciousness, Airedale NHS Trust made a request to withdraw feeding. 'E' was a 32-year-old woman with severe anorexia, who was under palliative care as she refused all nutrition. Her local authority sought judicial review from the Court of Protection to determine if she should be 'force-fed'.

Case law since *Bland* and *Re E* concerning the withdrawal of clinically assisted nutrition and hydration in patients with a PDC, and the refusal of feeding in anorexia cases, continues to raise legal and ethical questions.[7] This has focused primarily on a discussion of autonomy, rights, and the value of life. However, these issues will not be the focus of the discussion in this article. Instead, I will focus on the emotional responses and reasoning found in relevant areas of law. I will consider the extent to which emotion has influenced case law and legislation, and whether it has informed moral reasoning in this area, to develop a position on what role emotion *should* play in end-of-life decisions.

In the background of these practical questions is a more fundamental debate about the role of emotion in morality, and this is where the article begins. Both Kantian and utilitarian traditions in moral theory promote reasoning as the

[3] Natasja Raijmakers and others, 'Variation in attitudes towards artificial hydration at the end of life: a systematic literature review' (2011) 5(3) Current Opinion in Supportive & Palliative Care 265, 265.

[4] Steven Miles, 'Futile Feeding at the End of Life: Family Virtues and Treatment Decisions' (1987) 8 Theoretical Medicine 293, 295.

[5] *Airedale NHS Trust v Bland* [1993] AC 789 (HL), [1993] 2 WLR 316.

[6] *A Local Authority, E (by her Litigation Friend the Official Solicitor) v A Health Authority, E's Parents (also known as "Re E")* [2012] EWHC 1639 (COP), [2012] HRLR 29.

[7] *Bland* (n 5), *Re E* (n 6).

principle method of determining what we are morally required to do. These theories portray emotion as a negative influence on one's ability to reason, and do not consider it to be part of rational decision-making. However, I argue that a theory of emotion is essential to our understanding of not just how we respond to moral questions, but how we reason and make moral decisions. I posit that emotions are fundamental to how we develop our moral framework, and also provide the means by which we put moral reasoning into practice. Together with the legal analysis, this ethical analysis will be used to defend the view that there should be greater recognition of the crucial role emotion plays in end-of-life legal decision-making.

II. Emotions and Moral Reasoning

The role of emotion in moral reasoning is a long-standing point of contention. For instance, some traditional and ancient ethical theories advocated moral systems that are impartial and objective. Emotions are considered as a partiality that lead to irrational decision-making. It is still claimed by some philosophers, such as Peter Singer, that emotions corrupt rational moral reasoning, and should be controlled.[8] However, other philosophers have contested this claim. In this Part, I shall consider the claims against the role of emotions made by Kantian and utilitarian theorists. Singer, who has a utilitarian perspective, continues to disregard the role of emotion in moral reasoning. In response, the philosophical arguments which justify the role of emotion in moral reasoning, will be set out, in support of my position. I shall conclude that there is a central position for emotion in moral reasoning. This will provide the foundation for the legal analysis of cases involving feeding decisions at the end-of-life, which will be explored in Parts III to V.

A. Kantian ethics

Immanuel Kant discussed the role of emotion in his theory of morality. Kantian ethics takes a deontological approach. Deontology in a normative theory that states that morality is determined by the "rightness" of our actions, and not by their outcome.[9] The right choice conforms to moral norms and moral law.[10] Kant termed moral law the "categorical imperative"; which he claimed imposed duties on the agent.[11] He believed that conduct only had moral worth when it is

[8] Peter Singer, The Expanding Circle (Clarendon Press 1981) 93.

[9] Larry Alexander and Michael Moore, 'Deontological Ethics' (Stanford Encyclopedia of Philosophy, 17 October 2016) <https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/> accessed 20 August 2017.

[10] ibid.

[11] Julia Driver, Ethics: The Fundamentals (Blackwell Publishing 2007) 80.

driven by duty, in compliance with the categorical imperative. In his consideration of emotions, which he termed "inclinations", he claimed that though they are not necessarily a negative force, they have no moral worth.[12] Kant's views were in reaction to David Hume, who stated that "reason is, and ought only to be the slave of the passions, and can never pretend to be any other office than to serve and obey them".[13] Hume argued that "the passions" (or emotions) are the only motivation to human action. Reason can never motivate action, but is employed to reach the goals set by our emotions.[14] Kant refuted this as he believed inclinations to be too fickle as motivators of action. Reason, he believed, provided moral motivation and rational judgements about what we should do.[15] Kant went so far as to say that, when love is commanded out of duty, it is more morally significant than when it is motivated by a desire to do good.[16]

Stocker has argued that the traditional theorists such as Kant failed to appreciate the importance of motivational and emotional drivers of rational morality. Stocker argued that a life lead by duties and obligations is not fulfilling, nor does it evoke "moral goodness".[17] He posits that the conceptual disassociation of motivations from reasoned judgement is not applicable in the real world. He claims that to value something one must have a motive, and that "motive and reason must be in harmony for the values to be realised".[18] Stocker stated that this is a failure of not just Kant's deontological ethics, but also of utilitarianism and egoism.[19] These theories failed to consider value of the relationship between the subject and object of affection.[20] Stocker demonstrates this by considering love; he argues that without the commitment of an interdependent relationship, acting for the sake of duty will not be sufficient motivation to maintain the act.[21] The divide between the reason and motive becomes incomprehensible in real life situations and has pragmatic failings. Stocker goes so far as to say that without valuing the motivations, we dehumanise relationships, and fail to acknowledge what makes "a human life worth living".[22]

[12] ibid 86.
[13] David Hume, A Treatise of Human Nature (Dover Publications 2003) 295.
[14] Driver (n 11) 82.
[15] ibid 84.
[16] Immanuel Kant, Grounding for the Metaphysics of Morals on a Supposed Right to Lie because of Philanthropic Concerns (James Ellington tr, 3rd edn, Hackett 1981) 12.
[17] Michael Stocker, 'The Schizophrenia of Modern Ethical Theories' (1976) 73(14) The Journal of Philosophy 455.
[18] ibid.
[19] ibid 459.
[20] ibid.
[21] ibid 458.
[22] ibid 460.

B. Utilitarian ethics

Emotion in decision-making is fundamentally opposed to utilitarian ethics. Utilitarian ethics claim that the morally right decision is the one that maximises overall happiness and therefore, the greatest good is that which benefits the greatest number.[23] Those who argue in support of utilitarian ethics consider emotions to be a barrier to objective reasoning in determining the greatest good. Singer states that ethics evolved out of our capacity to reason.[24] Reasoning, he claims, is only acceptable when the agent is "disinterested" in their own or other's interests.[25] He states that one's interests are no more important than those of others, and that equal weight should be given to the interests of all.[26] Therefore, interests are partialities which should not motivate decision-making, as this will not be accepted as valid by other reasoned beings.[27] Personal relationships and emotions provide a typical example of a potential partiality, which should not influence the determination of the greatest good. He disapproves of charitable contributions motivated by emotional responses, and instead endorses a rational consideration of where one's donation would do the most good.[28] Though Singer has denied criticisms that he wants to divorce emotion from charitable behaviour, he still views it as a potentially corrupting influence. He claims that reason alone should determine the ethical direction of what to do.[29] Emotion leads to impulsive reactions, and if we are partial our actions will not be motivated to do the most good.[30]

The impartiality advocated by Singer has pragmatic failings, and raises questions as to whether we have greater moral obligations to those closest to us. For example, the decision to prioritise caring for one's own elderly parents, over other elderly and equally needy persons, is a result of the greater compassion one feels for their own parents. This clearly contradicts utilitarian principles, where "a decision must give equal weight to the interests of all affected by it".[31]

[23] Julia Driver, 'The History of Utilitarianism' (The Stanford Encyclopedia of Philosophy, 22 September 2014) <https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history/> accessed 2 August 2017.

[24] Singer (n 8) 111.

[25] ibid 93.

[26] ibid 111.

[27] ibid 93.

[28] Peter Singer, 'Precis: The Most Good You Can Do' (2016) 12(2) Journal of Global Ethics 132, 132.

[29] Anne Maclean, The Elimination of Morality: Reflections on Utilitarianism and Bioethics (Routledge 1993) 53–54.

[30] Peter Singer, 'Altruism and Emotion' The New York Times (New York, 10 December 2015). <www.nytimes.com/2015/12/11/opinion/peter-singer-on-altruism-and-emotion> accessed 2 August 2017.

[31] Singer, The Expanding Circle (n 8) 79.

Maclean has contested Singer's argument for impartiality. Instead, she submits that the recognition of special rights and obligations one has, such as those to one's own parents, is not a moral failing.[32] She states that avoiding partiality does not necessarily improve moral decision-making.[33] For example, it would seem morally wrong if one were to choose to donate all their money to a charity that helped needy people, and neglected the needs of their elderly parents. Maclean denounces Singer's separation of moral reasoning and human nature, stating that "there is no gap between the emotional and the rational components of human nature".[34] This argument poses that our emotional life is a precursor to rationality, rather than an obstacle.

Taylor's thesis sees emotional responses as normative, and connected to our evaluation of situations or events. Taylor disputes the claim that emotions are irrational, but goes further in saying that a deficiency of appropriate emotional responses is an indication of "moral failing or human short-coming".[35] An emotional reaction is justified when firstly the belief, of which the reaction is based, is well-founded; and secondly, that the reaction is appropriate to the situation. For example, if one saw a snake, believing it to be poisonous, it would be appropriate to feel fear. Furthermore, the reaction must be proportional to the stimulus. So, if one were to happen upon a poisonous snake in the wild, it might be appropriate to jump, scream, or run away. However, if one were to see a poisonous snake in a zoo, behind a glass screen, this response would be excessive, and therefore unjustified.[36] Taylor is not just saying that emotions are rational, but further, that their rationality is morally relevant.

The counter arguments posed by Stocker, Maclean and Taylor make a strong case for emotion as an essential component of our reasoning. They view the relationship between emotion, reasoning and morality, as much more co-dependant than the Kantian or utilitarian claims imply. Novel theories of morality have served to re-frame the debate, giving emotion a central role in our understanding of morality.

C. A new perspective: feminist ethics and the value of caring

Feminist ethics came about as a reaction to traditional theories, which were considered gender-biased, and dismissed typically female qualities as being morally

[32] Maclean (n 29) 59.
[33] ibid 61.
[34] ibid 69.
[35] Gabriele Taylor, 'Justifying the Emotions' (1975) 84(335) Mind 390, 390.
[36] ibid 392–393.

deficient.[37] Feminist philosophers disputed the entire framework of traditional theories which claimed women to be too emotional, personal, and "incapable of reason".[38] Female qualities assigned to women by essentialist theory further devalued and subordinated the woman's perspective. This extended to typically female 'activities' such as caring and mothering. Though these activities were praised, and even idealised, the undercurrent placed them as secondary to more 'important' male activities.[39] In response, novel moral systems have been formulated that advocate female qualities and activities, and incorporate the role of emotions in moral reasoning.

Fischer has written about the philosophical 'turn to affect' that occurred in the mid-1990s.[40] She describes the movement as a "paradigm shift in critical theorising", which brought the discussion of emotion and feeling to the forefront of the debate.[41] Progressive feminist theorists, such as Gilligan and Noddings, have posited an emotional, rather than the traditional rational, basis for morality.[42] Gilligan provided the basis for the care approach, and Noddings developed the normative ethical theory in her book *Caring.* Gilligan proposed that girls think differently to boys about moral issues and problems. She examined the work of Lawrence Kohlberg, and highlighted a fundamental gender-bias in his theory of moral development.[43] Kohlberg's understanding of morality was based on rules, principles, and justice; and derived from Kant and John Rawls.[44] He identified six stages of moral development, and formulated a test which he carried out on children. He found that girls tended to lag behind boys and claimed girls had a less developed sense of morality than boys.[45] Gilligan observed that though the boys tended to use logic or law to mediate situations, girls responded with communication through relationships. Gilligan disputed the claim that moral reasoning based on

[37] Jean Grimshaw, 'The Idea of Female Ethics' in Peter Singer (ed), A Companion to Ethics (Wiley 2013) 502.

[38] ibid.

[39] ibid.

[40] Clara Fischer, 'Feminist Philosophy, Pragmatism, and the "Turn to Affect": A Genealogical Critique' (2016) 31(4) Hypatia 810, 810.

[41] ibid 811.

[42] ibid 814.

[43] Carol Gilligan, In a Different Voice (Harvard University Press 1982) 100.

[44] Grimshaw (n 37) 503.

[45] Lawrence J Walker, 'Gender and Morality' in Melanie Killen and Judith Smetana (eds), Handbook of Moral Development (LEA 2006) 97.

relationships should be interpreted as being inferior to male patterns of moral reasoning.[46]

Noddings' ethical theory posed a new approach in moral reasoning. She saw traditional theories, such as Kantianism and utilitarianism, as having been discussed in "the language of the father: in principles and propositions, in terms such as justification, fairness and justice".[47] Instead she argued that the basis of human ethics is "caring".[48] She critiqued traditional theories for simplifying moral dilemmas to rules, which did not resemble real situations. When faced with a moral question, she posited that women typically need more information to come to a decision. Such as, the thoughts and feelings of all those involved. Female reasons, she stated, are based on "feelings, needs, impressions and a sense of personal ideal".[49] Which in turn had been judged unfairly as an inferior rationale than that of men.[50] Emotion is at the heart of her ethical theory and is based on the principle that all ethical behaviour stems from a universal "caring attitude".[51] She denied that justification is the principle measure of ethical conduct, advocating motivation instead. She refuted theories that claimed moral judgements could be tested in the same way as facts: that there is a definitive right or wrong. Instead she posed that moral judgements are not truths, but derive from a rational, caring attitude. Emotions in the context of a caring attitude are both responses and appraisals of a situation.[52] She further suggests that emotion does not necessarily require action to be complete; it can serve reflective purposes that restore the agent to a less stressful state.[53]

The feminist perspective has clearly laid out a place for emotion in reasoning and morality. The difficulty with removing emotion from rational thinking is hypothetical, as our emotional lives are intertwined with the cognitive processes required for reasoning. This poses a strong argument against the theories that claim: firstly, that emotions are irrational and illogical; and secondly, that emotions are not conducive to moral reasoning or appropriate moral action. In response to

46 Gilligan (n 43).
47 Nel Noddings, Caring: A Feminine Approach to Ethics & Moral Education (2nd edn, University of California Press 2013) 1.
48 ibid.
49 ibid 3.
50 ibid.
51 ibid 92.
52 ibid 142.
53 ibid 142–143.

these views emotion theorists have formed different approaches to explain how and why emotions influence our decision-making.

The moral decisions made in court can have the gravest consequences, and involve highly emotive situations. I argue that cases involving end-of-life feeding decisions are especially emotive. This is due to the concept of food being a basic and essential requirement. However, these concepts are tested when feeding is no longer sustaining but prolonging life, as is the case for patients with a PDC; or when food is viewed by the patient as detrimental to their existence, as is the case with anorexia. I have argued that there is a central role for emotion in moral reasoning. Therefore, the role of emotion in judicial reasoning in these cases will be particularly important. To explore this further, the legislation and case law will be set out and critiqued, to determine whether the decisions made in court are truly appreciative of the emotions; and further, to what extent do they influence judicial reasoning and the outcome of cases.

III. Prolonged Disorders Of Consciousness

Medical advances have developed treatments that allow recovery from even the most catastrophic brain damage; but this is not the case for everyone. Those who do not recover can be maintained in a state of perpetual limbo; suspended between life and death. The legislation and jurisdiction that apply to cases of PDC shall be described. Then, the specific emotions that have manifested in case law and its influence on legislation will be considered.

Disorders of consciousness (DCs) refer to states where a patient has wakefulness, but absent or reduced awareness of their surroundings. DCs result from brain injury where the brain-stem remains intact but areas of the cerebral cortex, which regulate awareness and higher-level brain functions, is profoundly damaged. DCs can be transient or permanent: they include coma, vegetative state and minimally conscious state. If a DC persists for more than four weeks it can be said to be a PDC and will be diagnosed as either a vegetative state or a minimally conscious state.[54]

In a vegetative state there is wakefulness with complete absence of environmental awareness, and in a minimally conscious state there is wakefulness with minimal awareness. A minimally conscious state presents with "inconsistent, but reproducible" behavioural evidence of awareness, which is greater than reflexive or automatic responses.[55] In 2015, the incidence of patients in a vegetative state

[54] Royal College of Physicians, Prolonged disorders of consciousness: National clinical guidelines (Royal College of Physicians 2015) 1.

[55] ibid 3.

in the UK was between 4,000 and 16,000; and three times as many were thought to be in minimally conscious state.[56] If consciousness has not been regained after one to two years the likelihood of regaining consciousness is very low, and those who do are likely to be left with profound disabilities.[57] The body is maintained with clinically assisted nutrition and hydration, and patients will require 24-hour specialist care and equipment to prevent complications.[58] In 2013, the cost of treating someone in a PDC was estimated at £7,500 per month.[59]

A. The law applied to prolonged disorders of consciousness

A patient with capacity has the right to refuse any medical treatment.[60] Medical law is built on the principle that treatment can only be given to a competent adult if they give their consent. To provide treatment where the risks outweigh the benefits is malpractice.[61] However, those in a PDC lack the mental capacity to make a decision and their wishes are scarcely known. For those lacking capacity, we need to consider the Mental Capacity Act 2005 (MCA 2005) which came into force in 2007. It serves as the legal framework to determine capacity and provides safeguards such as the patient's best interests checklist, advance decisions, and creating a lasting power of attorney.

B. Advance decisions

An Advanced Decision to Refuse Treatment (ADRT) can be made by a competent adult over 18 to specify what treatments they would accept or refuse if they were to lose decision-making capacity.[62] An ADRT can be cancelled or amended by the competent individual at any time.[63] If the ADRT refuses life-sustaining treatment it must be in writing, signed and witnessed, and state that the

[56] Sarah Bunn and Zoë Fritz, 'Vegetative and Minimally Conscious States' (Houses of Parliament, Parliamentary Office of Science & Technology 2015) 1 <https://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-489#fullreport> accessed 2 August 2017.

[57] Royal College of Physicians (n 54) 9.

[58] ibid 3.

[59] ibid.

[60] Department for Constitutional Affairs, 'Mental Capacity Act 2005 Code of Practice: Issued by the Lord Chancellor on 23 April 2007 in accordance with sections 42 and 43 of the Act' (published by The Stationery Office on behalf of the Department for Constitutional Affairs 2007) 20 <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/497253/Mental-capacity-act-code-of-practice.pdf> accessed 2 August 2017.

[61] *F v West Berkshire HA* [1991] UKHL 1, 1.

[62] Mental Capacity Act 2005 Code of Practice (n 60) 159.

[63] ibid.

person is aware that the consequences, if actioned, could be life-threatening.[64] "A valid and applicable advanced decision to refuse treatment has the same force as a contemporaneous decision", and when applicable it must be followed by clinicians or they could face criminal prosecution or civil liability.[65]

If an ADRT has not been written correctly, or is not applicable to the decision that needs to be made, it is not binding but may be considered as part of a best interests decision. In *Re D*, 'D' had put his wishes into a letter stating that he refused invasive treatments that would prolong a reduced quality of life.[66] D was in vegetative state, but as his signature had not been witnessed the ADRT was not valid. Nonetheless, the court came to the same decision in his best interests, and clinically assisted nutrition and hydration was withdrawn. This highlights an issue with ADRT: without legal advice, the layperson may be unaware of the stipulations to producing a legitimate ADRT. It will also be invalid if it was withdrawn by the individual before they lost capacity; if it does not specify what treatment they are refusing; if there is conflict with a subsequent ADRT or a lasting power of attorney; or if there are reasonable grounds to believe that the ADRT no longer represents the patient's decision.[67] On this final point, it has been applied to circumstances where there have been advances in medications or treatments since the time the ADRT was written up. In these circumstances, it is deemed that the patient would have been unaware of the anticipated better outcome; therefore, the ADRT is no longer deemed valid.[68] Secondly, this applies where there is reasonable belief that the ADRT no longer represents the patient's beliefs or values. In *HE v A Hospital Trust;* 'AE' was a 24-year-old woman who was born Muslim before becoming a Jehovah's Witness and writing an advance decision to refuse blood transfusions. Later, was engaged to a Muslim man and stopped attending Jehovah's Witness meetings and services. When she lost capacity, her father sought a declaration from the court that the ADRT was no longer applicable. Munby J granted this after stating that decisions should favour the preservation of life when there were doubts to the validity of an ADRT.[69]

There are several practical difficulties in using an ADRT to protect one's wishes. First, to ensure its validity, legal advice will be required when writing it up. Secondly, medical advice may be desirable to understand the range of treatments and circumstances one might find themselves in when incapacitated, to ensure the ADRT is suitably specific. Thirdly, once produced it needs to be proven to be up-to-date or revised when wishes or circumstances change. Finally, one needs to

[64] ibid.
[65] ibid para 9.1, 9.2.
[66] *Re D; An NHS Trust v D* [2012] EWHC 886 (COP) [15].
[67] Explanatory notes to the MCA 2005, s 25.
[68] ibid, para 88.
[69] *HE v A Hospital Trust* [2003] EWHC 1017 (Fam) [49], [50].

put in place a process for ensuring that if one loses capacity the ADRT will be shown to the decision-makers. If all these difficulties are accounted for, the ADRT will provide a robust safeguard. However, the barriers are apparent, in 2015 it was reported that only 4% of the population produced an ADRT.[70] The MCA 2005 provides an alternative safeguard: appointing a proxy decision-maker, with the hope that some of these obstacles would be overcome.

C. Proxy decision-makers

The only form of proxy decision-maker who, if authorised, can make decisions to withdraw or withhold life-sustaining treatment on behalf of an incapacitated adult is a lasting power of attorney. This power, introduced by the MCA 2005, allows a patient to grant a person lasting power of attorney (hereinafter referred to as the "LPA") to make medical decisions on their behalf, if they were to lose capacity in the future.[71] The power can be exercised to make the decision to withdraw life-sustaining treatment if the patient specifically expressed this in the documentation.[72] The patient can appoint one or more LPAs, who either work jointly or severally.[73] The LPA's powers are restricted in that they cannot override an applicable and valid ADRT, nor can they make a decision that is not deemed to be in the patient's 'best interests'.[74] This is where the LPA's powers do not equate to a competent person's refusal. The LPA still has to apply the best interests checklist when making decisions and must not be motivated in any way by the desire to bring about the donor's death.[75] This is contentious: the patient may appoint an LPA to act on their behalf in unforeseen circumstances. Yet the LPA's influence may be no greater than any other next-of-kin in the medical team's determination of a best interests decision. As with the legal conditions required of an ADRT, the consequences of the MCA 2005 stipulations may not be fully understood by the patient when appointing an LPA. Therefore, they may also find themselves receiving treatment they would never have consented to.

D. Best interests

If the patient does not have an applicable and valid ADRT and they lose capacity, the medical team must make a best interests decision on their behalf.[76] The

[70] Bunn and Fritz (n 56) 3.
[71] Mental Capacity Act 2005, s 9(1)(a).
[72] Explanatory Notes to the Mental Capacity Act 2005, para 58.
[73] Mental Capacity Act 2005, s 10(4).
[74] Mental Capacity Act 2005 Code of Practice (n 60) para 7.24.
[75] Mental Capacity Act 2005, s 4(8)(a).
[76] ibid s 4.

points for consideration in determining a best interests decision are listed in section 4 of the MCA 2005. This includes ensuring that no presumptions are made about what the patient would think to be in their best interests;[77] and that the views of those close to the patient, such as the next-of-kin or LPA, are taken into account.[78] The patient's wishes, feelings, beliefs, and values, must also be ascertained where possible.[79] In decisions relating to life-sustaining treatment, the decision-maker must not "be motivated by a desire to bring about his death".[80] There are, however, difficulties and uncertainties about how the test should be applied: namely, there is no indication of what weight should be given to the different factors, and this is left to the decision-maker to determine. Though it allows flexibility in its application, it may also lead to inconsistency and insufficient regard for the patient's wishes and feelings. This has led to calls to reform the best interest test to comply with the UN Convention on the Rights of Persons with Disabilities (UNCRPD), which was ratified by the UK in 2009.[81] Though recent case law is placing greater emphasis the patient's wishes and feelings in determining best interests, though this has yet to be formalised.[82]

The determination of best interests starts with the assumption that life should continue. However, when treatment is deemed "futile, overly burdensome or intolerable for the patient or where there is no prospect of recovery", the decision to withdraw treatment is deemed justifiably in the patient's best interests.[83] This follows from Lord Keith's statement in *Bland* that "existence in a vegetative state with no prospect of recovery is by that opinion regarded as not being a benefit, and that, if not unarguably correct, at least forms a proper basis for the decision to discontinue treatment and care".[84] It is therefore difficult to reason how continuation of treatment in this instance could ever be viewed as being in the patient's best interests. If the diagnosis is indicative of no prospect of recovery, then the continued treatment could be argued to be at least malpractice, or at worst assault. The court's decision therefore hinges on the diagnosis. The first request to withdraw clinically assisted nutrition and hydration from a patient with a minimally conscious state diagnosis was in *W v M*,[85] but the application was

[77] ibid s 4(1)(b).
[78] ibid s 4(7).
[79] ibid s 4(6).
[80] ibid s 4(5).
[81] Law Commission, Mental Capacity and Deprivation of Liberty (Law Com No. 372, 2017) 161.
[82] *Briggs v Briggs* [2016] EWCOP 53, [2017] 4 WLR 37 [7].
[83] *Re C* [2010] EWHC 3448 (COP) [59].
[84] *Bland* (n 5) 859.
[85] *W v M* [2011] EWHC 2443 (Fam), [2012] 1 WLR 1653.

not granted as withdrawal was not deemed to be in her best interests. Newton J came to a similar conclusion in *St George's Healthcare NHS Trust v P&Q*, but in this case the determination of the patient's wishes and feelings formed a significant role in the ruling.[86] the patient's religious beliefs, his family's disagreement with the medical team, and his diagnosis favoured the continuation of life-sustaining treatment. In determining best interests, the court's role has historically been one of scrutinising the diagnosis and acting accordingly. However, as greater weight is placed on wishes, feelings, beliefs and values, the emphasis falls on a determination how the patient would view that decision.[87] This has marked the most significant change in the legal jurisprudence in these cases: the removal of the requirement for judicial approval when withdrawal is unanimously deemed to be in the patient's best interests.

E. Judicial approval

2017 saw three landmark cases that questioned and held that there is no longer a requirement for clinicians to seek court approval, when there is no dispute that withdrawal is in the patient's best interests.[88] Up until last year, approval from the Court of Protection was advised in all incidences.[89] This specification had been heavily criticised, as it incurred significant time delays and costs.[90] The requirement originated from the recommendation made by Lord Goff in 1993: in *Bland*, he stated that judicial approval would be "desirable" for the reassurance of families and the public.[91] However, he also stated that once a "body of experience and practice has been built up" the need for an application would not be necessary in every case.[92] Yet this has taken two and a half decades to be realised, notably

[86] *St George's Healthcare NHS Trust v P&Q* [2015] EWCOP 42.

[87] British Medical Association, Decisions to withdraw clinically-assisted nutrition and hydration (CANH) from patients in permanent vegetative state (PVS) or minimally conscious state (MCS) following sudden-onset profound brain injury: Interim guidance for health professionals in England and Wales (British Medical Association 2017) 2.

[88] *NHS Trust v Mr Y and Mrs Y* [2017] EWHC 2866 (QB), *Director of Legal Aid Casework & Ors v Briggs* [2017] EWCA Civ 1169, *M v A Hospital* [2017] EWCOP 19.

[89] Court of Protection, Practice Direction 9E: Applications relating to serious medical treatment (Courts and Tribunals Judiciary 2015) para 5(a).

[90] Adam Formby, Richard Cookson and Simon Halliday, 'Cost Analysis of the Legal Declaratory Relief Requirement for Withdrawing Clinically Assisted Nutrition and Hydration (CANH) from Patients in the Permanent Vegetative State (PVS) in England and Wales' (2015) CHE Research Paper 108, iii <https://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP108_cost_analysis_CANH_PVS_declaratory_relief.pdf> accessed 3 August 2017.

[91] *Bland* (n 5) 815–816.

[92] ibid.

because the recommendation was established in legislation, such as the MCA 2005 Code of Practice and Court of Protection Rules.[93] The British Medical Association guidance acknowledged that this instruction was not intended to be immutable, and hoped that court review would not be required indefinitely.[94] They noted that there is often little practical difference between clinically assisted nutrition and hydration withdrawal in these circumstances, and other equally serious situations where it is not seen as a benefit.[95] The original purpose of judicial approval was to reassure families, yet research found the effect to be the contrary.[96]

These changes will certainly be welcomed by many. In recent years there has been mounting support for reform of the law,[97] though others may call to question why this reform took so long to come, and why so many judges were unwilling to question its necessity. It has been commented that the judges' language in these cases perpetually reinforced it as "a decision for the courts".[98] Lord Goff's recommendation undoubtedly was an attempt to quell concerns that the ruling would lead to euthanasia by the back door. Yet, I also posit that the nature of feeding as an act of providing nourishment, and the abhorrence of starvation as a means of death, also explains why reform took so long to come.

F. Feelings about life and feelings about food

There are two principle issues that provoke emotional response in PDC cases: these are the feelings one has about what it is to have a meaningful life, and concerns about imposing one's own standards on the incapacitated. The other is the emotion roused by the idea of withdrawing nutrition and hydration. The emotional response and reasoning of these issues are complex and may vary depending on the subject's relation to the patient. The discussion here will focus on

[93] Mental Capacity Act 2005 Code of Practice (n 60) para 8.18; The Court of Protection Rules 2017, SI 2017/1035.

[94] British Medical Association, Withholding and Withdrawing Life Prolonging Medical Treatment: Guidance for Decision Making (3rd edn, Blackwell Publishing 2007) para 30.2.

[95] ibid.

[96] Celia Kitzinger and Jenny Kitzinger, 'Court Applications for Withdrawal of Artificial Nutrition and Hydration from Patients in a Permanent Vegetative State: Family Experiences' (2016) 42(1) Journal of Medical Ethics 11.

[97] Simon Halliday, Adam Formby and Richard Cookson, 'An Assessment of the Court's Role in the Withdrawal of Clinically Assisted Nutrition and Hydration from Patients in the Permanent Vegetative State' (2015) 23(4) Med L Rev 556, 587.

[98] Richard Huxtable and Giles Birchley, 'Seeking Certainty? Judicial Approaches to the (Non-)Treatment of Minimally Conscious Patients' (2017) 25(3) Med L Rev 428, 438.

the latter, but there is a complex interaction between the two, and it is not possible to consider one wholly in isolation.

The greatest emotional burden in these cases falls on the relatives. Kitzinger and Kitzinger found that two thirds of relatives believed that the patient would rather be dead than sustained in their condition.[99] Feelings about the use of feeding tubes are mixed. Some see it as a non-negotiable basic act of nurturance, whereas others consider it to be an "unnatural act of heroic medicine and a technological means of denying death".[100] Nonetheless, even those who believe the patient would not want to be alive, find the prospect of allowing starvation too cruel and barbaric to even contemplate.[101] Families long for a peaceful death where nature takes its course; and death by withdrawal of clinically assisted nutrition and hydration is often not perceived this way.[102] Kitzinger and Kitzinger reflect that "failing to feed… is a highly emotive issue with deep cultural resonance".[103] There are two reasons why death by starvation is viewed this way. First, the death is slow, it can take three weeks or longer. The prospect of watching their loved one waste away for that time is often unbearable. It also provides a long period of time to regret and reconsider their decision. Secondly, starvation is considered a 'painful' death. The assurance that they can no longer feel pain, or will be provided with pain relief, does little to alleviate their concern. Everyone has experienced hunger pains, and is equally aware of how easily they are remedied after eating. Ian Miller has commented that western societies find the idea of hunger particularly "unacceptable".[104] The difference in court rulings for those in a vegetative state and minimally conscious state may be influenced by the idea of starvation pain. The decision to deny withdrawal in *W v M*, who had a minimally conscious state diagnosis, could be viewed as a reaction to the abhorrence of starvation.[105] It could be argued that this decision was influenced by the patient's increased consciousness and therefore perceived ability to feel pain, with the reasoning that starvation is unacceptable for those able to experience it.

The emotional response to starvation is not just due to the means of death, it is also related to the carer's perception of their caregiving responsibilities. In *Re C,* the staff at the unit opposed the application for withdrawal because it "is against

[99] Kitzinger and Kitzinger (n 96) 158.

[100] Sharon Kaufman, And a Time to Die: How American Hospitals Shape the End of Life (University of Chicago Press 2006) 188.

[101] Kitzinger and Kitzinger (n 96) 159.

[102] ibid.

[103] ibid.

[104] Ian Miller, 'Starving to death in medical care: Ethics, food, emotions and dying in Britain and America, 1970s–1990s' (2017) 12(1) BioSocieties 89, 93.

[105] *W v M* (n 85).

the unit's philosophy of care".[106] The nature of feeding in providing nurture and care is indisputable. Erde and Herring commented on nursing resistance to the introduction of pump-feeding to nursing homes in the 1980s (the pump provided a continuous feed, eliminating the need for nurses to provide bolus feeds by syringe).[107] Nurses equated the loss of feeding time with a "loss of nurturing time".[108] For healthcare professionals the idea of allowing preventable suffering is particularly intolerable, and goes against caring responsibilities. Not only does it oppose the staff's perception of their own duties, but there is societal opposition as it "clashes with expectations of medical care".[109]

The ruling in *Bland* has not alleviated the emotional dilemma faced by families and clinicians.[110] Though Hoffmann J made note of the "emotional symbolism of food" and its power to evoke "deeply intuitive feelings".[111] The redefinition of clinically assisted nutrition and hydration to "medical treatment" was an attempt to reclassify his death as "letting nature take its course".[112] However, as we have seen in the Kitzinger and Kitzinger study, it is not perceived to be doing anything of the sort. Though the court goes to great lengths to try to sympathise with the emotions of the family and clinicians, such as emphasising the tragic nature of the case, it has failed to appreciate how the means of death will impact the family's reasoning.[113] Discontinuing feeding is not perceived as a natural death. This is demonstrated in the words of Karen Quinlan's father (a prominent case in the United States), who when asked if they should stop clinically assisted nutrition and hydration as well as her ventilation said: "oh no, that is her nourishment".[114] The focus of withdrawal in these cases being an omission, rather than an act, is juxtaposed with the emotional experience of those involved. This accounts for why, when judicial approval was required, only a tiny proportion of cases are ever brought to court. Kitzinger and Kitzinger's study found that families were frustrated that the courts had put them in the position where withdrawing clinically assisted nutrition and hydration was the only "legal exit route", leading some to even consider murder as a better alternative.[115] The court concerns itself with trying to be objective and

[106] Re C (n 83) [31].
[107] Edmund Erde and Marvin Herring, 'A discussion of some moral issues in nutrition and feeding' (1985) Journal of Medical Humanities and Bioethics 5, 6.
[108] ibid 7.
[109] Miller (n 104) 90.
[110] *Bland* (n 5).
[111] ibid 832.
[112] ibid 445.
[113] Huxtable and Birchley (n 98) 435.
[114] Daniel Callahan, 'On Feeding the Dying' (1983) 13(5) The Hastings Center Report 22, 22.
[115] Kitzinger and Kitzinger (n 96) 160.

impartial, and judicial approval was a defence to avoid accusations of euthanasia at all costs. Though the issue that causes the greater emotional pain is that the only option for families is to allow their relative to starve to death. This has not been sufficiently discussed in the courts and will continue to be a principle cause of distress for relatives and clinicians.

There are specific issues that arise in PDC cases due to the minimal prospect of recovery and the patient's inability to consent; but these are not the issues for those with anorexia. Here, the patient can usually fully participate in the discussions, and their refusal of food is often seen as the only barrier to a full recovery. In anorexia cases, food remains central to the legal and ethical dilemmas, and provokes strong emotional reaction. For those in a PDC the question has been whether to withdraw feeding, for those with anorexia it is whether to forcibly impose feeding. To consider the role of emotion in moral reasoning, these different cases need to be described and compared, this will be tackled in Part V of this article. Next, the legislation and case law applicable to patients with anorexia will be described and appraised.

IV. Anorexia Nervosa

Anorexia is an eating disorder characterised by low body weight. Persons with anorexia control their weight by restricting calorie intake and the types of foods eaten. Sufferers of anorexia may also exercise excessively, purge by vomiting or using laxatives, or binge eat.[116] The physiological effects of starvation cause the body to become emaciated and every major organ system can be compromised.[117] Anorexia has one of the highest mortality rates of any psychiatric condition.[118] Treatment for anorexia combines psychotherapies with monitored weight gain. The National Institute of Clinical Excellence (NICE) recommend that people with anorexia should be treated in the community where possible; however, those with severe physical complications may require inpatient treatment.[119] The intense fear of weight gain, as well as the distorted perception of their own weight, may result in non-compliance or avoidance of treatment services. Treatment may then need to be carried out against the individual's wishes if their weight is dangerously low. It is in these circumstances that the courts have had to intervene to determine

[116] National Economic and Development Authority, 'Anorexia: Overview and Statistics' <https://www.nationaleatingdisorders.org/anorexia-nervosa> accessed 3 August 2017.

[117] American Psychiatric Association (APA), Diagnostic and Statistical Manual of Mental Disorders: DSM–5 (5th edn, APA 2013) 341.

[118] Royal College of Psychiatrists, MARSIPAN: Management of Really Sick Patients with Anorexia Nervosa (2nd edn, Royal College of Psychiatrists 2014) 8.

[119] National Institute of Clinical Excellence, Eating disorders: recognition and treatment (NG69) (National Institute of Clinical Excellence 2017) paras 1.11.1, 1.11.3.

whether it is right to impose compulsory treatment in the form of 'force-feeding'. Force-feeding is typically administered through a nasogastric tube: a fine tube that is inserted into the nose, down the back of the throat and into the stomach. Liquid feed, fluids and medications can then be passed through it. Inserting the tube can be uncomfortable for the patient, particularly if they are resisting, but if the nasogastric tube is in the correct position and left undisturbed it should be quite comfortable.

A. The law applied to anorexia

A person with anorexia can be detained in hospital for treatment under the Mental Health Act 1983 (as amended in 2007) (MHA 1983), irrespective of whether they consent if the following three conditions are met.[120] First, the MHA 1983 can only be applied to treat 'mental disorders', which includes eating disorders.[121] A mental disorder is "any disorder or disability of the mind". [122] The European Convention on Human Rights (ECHR) uses the term "persons of unsound mind". Both legislations permit a broad interpretation and do not precisely define what this includes.[123] ECHR guidance states that some flexibility should be permitted in its interpretation because "psychiatry is an evolving field, both medically and in social attitudes".[124] This allows inclusion criteria to be adaptive, however it also risks being discriminative. In anorexia, this has been shown to implicate gender bias. It is portrayed as a female disorder, to the disadvantage of male sufferers. Assessments have been found to underscore men resulting in under-diagnosis. [125] Attempts to re-balance these issues include the use of more sex-neutral diagnostic

[120] Mental Health Act 1983, s 63.

[121] Department of Health, 'Mental Health Act 1983: Code of Practice: Presented to Parliament pursuant to section 118 of the Mental Health Act 1983' (published by The Stationery Office on behalf of the Department of Health 2015) 26 <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/497253/Mental-capacity-act-code-of-practice.pdf> accessed 2 August 2017.

[122] Mental Health Act 1983, s 1(1).

[123] European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, Article 5(1)(e).

[124] ECtHR, Guide on Article 5 of the Convention: Right to Liberty and Security (ECtHR, 2014) para 87.

[125] Alison Darcy and Iris Lin, 'Are we asking the right questions? A review of assessment of males with eating disorders' (2012) 20(5) Eating Disorders: The Journal of Treatment and Prevention 416, 417.

criteria in the DSM–5.[126] Nevertheless, there remain potential dangers in using social attitudes as a determining factor.

Secondly, the individual's health and safety must be sufficiently at risk to warrant hospital detention and treatment.[127] This could include a risk of suicide, or the physical effects of starvation and malnutrition being life-threatening. This is supported by NICE who advocate refeeding when physical health is severely compromised.[128]

Finally, appropriate treatment must be available. The administration of force-feeding as treatment for anorexia is controversial. Food is not normally considered to be medicine, however in *Bland*, it was agreed that clinically assisted nutrition and hydration is to be considered medical treatment and not basic care.[129] Treatment has since been broadened to include symptoms or manifestations that stem from the mental disorder.[130] *Riverside Mental Health NHS Trust v Fox* was the first anorexia case to authorise physical or pharmacological restraint as part of the treatment required to apply force-feeding.[131] The judge reasoned this by stating that in order to access therapy for her eating disorder she first needed to gain weight, and he concluded that force-feeding was therefore treatment for anorexia under section 145.[132] In *Re KB*, of the same year, the court held that NGT feeding was treatment for mental disorder for those with anorexia. The judge stated that "relieving symptoms is just as much a part of treatment as relieving the underlying cause".[133]

The MHA 1983 is typically required if the individual is competent but refusing treatment. The MHA 1983 would not be required if the patient were competent and consenting to treatment; in which case, the normal principles of medical law would apply. For a patient who lacked capacity, they could be treated in their best interests under the MCA 2005.[134] However, due to anorexia being termed a mental disorder, and the patient's likely refusal of treatment, the MHA 1983 is often applied. Furthermore, the courts have consistently ruled that persons

[126] APA (n 117) 338–345.
[127] Mental Health Act 1983, s 3(2).
[128] National Institute of Clinical Excellence (n 119) para 1.11.1.
[129] *Bland* (n 5).
[130] Mental Health Act 1983 Code of Practice (n 121) para 24.4.
[131] *Riverside Mental Health NHS Trust v Fox* [1994] 1 FLR 614.
[132] Mental Health Act 1983, s 145(4).
[133] *Re KB (adult) (mental patient: medical treatment)* [1994] 19 BMLR 144, 146.
[134] Jonathan Herring, Medical Law and Ethics (6th Edition, OUP 2016) 580.

with anorexia lack capacity. This is a contentious area of debate, as is the concept of force-feeding being in best interests.

B. Capacity

Adults with capacity, including those with mental illness, have the right to refuse any medical treatment for whatever reason, even if without it they will die.[135] The test of competence set out in *Re C*, formed the basis of the capacity assessment in the MCA 2005.[136] Capacity is determined using a functional and diagnostic test. The functional test assesses an individual's ability to understand, retain, use or weigh the relevant information and to communicate a decision.[137] The diagnostic test states that incapacity can only occur if it can be proven that there is an impairment or disturbance to the functioning of the mind or brain.[138] Capacity should be assumed unless proven otherwise.[139] Nevertheless, for those with anorexia, the courts have consistently ruled that the patient lacks the capacity to make a decision about refusing food. The patient's reasoning fails as they are not deemed to *believe* the information around risks.

The courts consistency is persistent: in *Re E*, Jackson J stated "E's obsessive fear of weight gain makes her unable of weighing the advantages and disadvantages of eating in any meaningful way".[140] In *Re L*, King J found that Ms L understood that she was close to death "but she has no deep understanding of her position".[141] Similarly, the two expert assessments of Ms X agreed that she was unable to use and weigh up the information, as she was "unable to apply the information to herself or believe in the need for it"; Dr Glover went further by doubting Ms X's ability to understand or retain all the relevant information.[142] In another recent case, Hayden J said Z "never fully comprehended the consequences of her behaviour in relation to food or nutrition".[143] Though the diagnosis of a particular mental illness should not presume incapacity, anorexia could be argued as an exception to the rule.[144]

Anorexia is invariably deemed to impair the patient's ability to use or understand information around food, regardless of how eloquently the patient articulates their preferences or understanding. This premise has long persisted in

[135] *Re T (Adult: Refusal of Medical Treatment)* [1992] EWCA Civ 18 [3].
[136] *Re C (Adult: Refusal of Treatment)* [1994] 1 WLR 290.
[137] Mental Capacity Act 2005, s 3(1).
[138] ibid s 2.
[139] ibid s 1(2).
[140] *Re E* (n 6) [49].
[141] *The NHS Trust v L* [2012] EWHC 2741 (COP) [49].
[142] *A Foundation Trust v Ms X* [2014] EWCOP 35 [27], [28].
[143] *Cheshire & Wirral Partnership NHS Foundation Trust v Z* [2016] EWCOP 56 [5].
[144] Mental Capacity Act 2005 Code of Practice (n 60) para 4.48.

medical and legal thinking.[145] The 'absolute presumption' of incapacity has been criticised for only interpreting the decision to be in regards to nutrition, rather than the patient's wider understanding of their quality of life.[146] Frequently those with anorexia are deemed to have capacity regarding other areas of their life; for example Ms X could make decisions in regards to her alcoholism, and Ms L could consent to or refuse antibiotics for her pneumonia.[147] Though guidance states capacity assessments are decision specific, there also seems to be no consideration that the patient will be able to regain capacity in the future.[148] Munby J stated that believing information was essential in order to comprehend, understand and weigh that information.[149] However, the requirement of a 'belief component' has been contested: Coggon has argued that one does not need to believe in something to understand it, and has called for the removal of the belief requirement from mental capacity law.[150]

There are several reasons why the anorexic's lack of capacity is contentious. Firstly, there would be repercussions if a court deemed the patient to have capacity. This should not influence the judge's decision, but it is important to recognise the potential for bias. The capacitous patient has an absolute right to refuse medical treatment, and a court would find it very difficult to oppose a competent refusal. If the patient was found to have capacity this could raise concerns of the legality of the compulsory treatment carried out not just to that patient, but to other anorexia sufferers receiving compulsory treatment. Secondly, the MHA 1983 is a particularly paternal piece of legislation, and for those with anorexia this seems acutely apparent. All court cases have involved young women, many of whom developed anorexia before puberty, and have spent large parts of their lives in hospitals. In *Re E* (in which Jackson J ruled in favour of force-feeding); E's abilities and intelligence are mentioned multiple times.[151] Yet, these are not given as reasons to respect her autonomous decision, but a rationale for why she must be 'saved'. Lastly, there seems to be a strong reluctance to find the anorexic's decision unwise, but capacitous. I believe this is partly due to a misperception of what food and nutrition mean to different people. For those with a 'healthy' relationship to food; it not only provides sustenance, but also enjoyment and positive social interactions.

[145] Ian Kennedy and Andrew Grubb, Medical Law: Text and Materials (2nd edn, Butterworths Law 1994) 143–144.

[146] Daniel Wang, 'Mental Capacity Act, Anorexia Nervosa and the Choice Between Life-Prolonging Treatment and Palliative Care: *A NHS Foundation Trust v Ms X*' (2015) 78(5) Med L Rev 871, 871.

[147] *Ms X* (n 142) [30], and *Re L* (n 141) [54].

[148] Mental Capacity Act 2005 Code of Practice (n 60) para 4.4.

[149] *X v MM and KM* [2007] EWHC 2003 (Fam) [81].

[150] John Coggon, 'Alcohol dependence and anorexia nervosa: Individual autonomy and the jurisdiction of the Court of Protection' (2015) 23(4) Med L Rev 659, 667.

[151] *Re E* (n 6) [5], [16], [75], [101], [132].

To question this fundamental truth for 'healthy' eaters appears to be so jarring that it cannot be conceived to be rational to think otherwise. For whatever the reason, the anorexic is yet to be found to have the capacity to decide for themselves, and it resides with others to make a decision on their behalf.

C. Best interests

The determination of best interests in anorexia cases has shown a marked change in judicial thinking. Force-feeding patients with anorexia has long been criticised by medical and legal commentators, yet it was only comparatively recently that the courts have become reluctant to impose invasive treatments. Force-feeding aims to relieve the physical symptoms that stem from the mental disorder. However, the irrevocable long-term damage caused by taking away control, "the anorexic's holy grail",[152] for short-term gain is considered by many to be contradictory. The landmark ruling in *Re L* was significant in that it handed power back to the anorexia sufferer, and allowed Ms L to decide for herself whether she would accept nutritional support.[153]

The legal jurisprudence in determining best interests in anorexia is similar to that described for those in a PDC.[154] However, the apparent difference is that those with anorexia can participate in the decision-making, yet this seems to create other difficulties for the courts. Courts are under increasing pressure to keep the person at the centre of the process as evidence shows that the patient's involvement improves outcomes.[155] Court of Protection judges have been motivated to meet the patient prior to the trial to gain a more holistic understanding of the person.[156] The MCA 2005 acknowledges the patient by requiring the consideration of past and present wishes, feelings, beliefs, and values.[157] The difficulty in applying the MCA 2005 to those with anorexia is that their primary and consistent wish is to be left alone. Yet the lack of hierarchy in determining the weight to be accorded to the patient's wishes and feelings has been criticised for being "unhelpfully vague".[158] Their wishes seem juxtaposed to the presumption that we should all value our

[152] Penney Lewis, 'Feeding Anorexic Patients Who Refuse Food' (1999) 7(1) Med L Rev 21, 32.

[153] *The NHS Trust v L* (n 141) [71].

[154] Mental Capacity Act 2005, s 4.

[155] Val Williams and others, Making Best Interests Decisions: People and Processes (Mental Health Foundation 2012) 83 <www.mentalhealth.org.uk/sites/default/files/BIDS_report_24-02-12_FINAL1.pdf> accessed 4 August 2017.

[156] See, for example, *CC v KK and STCC* [2012] EWCOP 2136, and *Wye Valley NHS Trust v B* [2015] EWCOP 60.

[157] Mental Capacity Act 2005, s 4(6)(a)(b).

[158] Nell Munro, 'Taking wishes and feelings seriously: the views of people lacking capacity in Court of Protection decision-making' (2014) 36(1) Journal of Social Welfare & Family Law 59, 62.

own lives. Clough has commented that there is a "persistence of value judgements about the agency of the person with anorexia".[159] Instead the courts take a "narrow, biomedical view" that often leans heavily on the opinion of the medical experts.[160] Clough goes on to say that even though the change in judicial approach could be interpreted as a greater understanding of the patient's will and preferences; she states that in neither *Re L* nor *Ms X* were the complexity of their views satisfactorily explored.[161] Munro criticised the judge in *Re E* for making no effort to engage with the subtleties and contradictions of E's views.[162]

D. A plea for control: the emotional response and reasoning

The central issue—forcing treatment upon someone who vehemently does not want to receive it—provokes intense emotional response. The emotions of the patient, the family, and clinicians may clash, and precede opposing interpretations of best interests. The involvement of the court at this time introduces a further stress for those involved, as well as the judge's own feelings and emotions. In this section, I shall explore the trends in the emotional responses of different participants. I will consider how this played into their reasoning and what influence this may have had on the outcome of the trial.

When treating anorexia clinicians must appreciate the patient's tumultuous emotions: "appalling despair, disgust, upset, sadness about what one has 'done to oneself'".[163] Those cases that end up in court are the most serious, with the most persistent anorexia. Therapy in these cases has not been able to reconcile the deep-rooted feelings and emotions that underpin the patient's reasoning and emotional life. The control of nutritional intake has been the patient's means of managing their emotions. Lewis has observed that "force-feeding crushes the patient's will, destroying who the patient is".[164] This is demonstrated in the case law, such as Ms X's anorexia and alcoholism being described as "the very essence of her life".[165] The loss of identity was apparent in *Re E*, who 'sees her life as pointless'. The anorexic's reaction to their distress has been a plea to retain control over their own lives. Ms X "has repeatedly requested that we do not detain or forcibly feed her" and in expressing her own views in a letter stated that she had 'had enough of the

[159] Beverley Clough, 'Anorexia, Capacity and Best Interests: Developments in the Court of Protection since the Mental Capacity Act 2005' (2016) 24(3) Med L Rev 434, 439.
[160] ibid 439, 444.
[161] ibid 443.
[162] Munro (n 158) 66.
[163] Susie Orbach, Hunger Strike (Faber and Faber 1986) 139.
[164] Lewis (n 152) 33.
[165] *Ms X* (n 142) [21].

continual pressure of mental health staff and services'.[166] Ms W similarly wrote "I have no control, things I would like I am being denied", and E "wants to be allowed to make her own choices".[167]

The anorexic rationalises her plea to be left alone from two perspectives. Firstly, that force-feeding is futile. It is difficult to contest the anorexic's perspective in this instance: in each case there have been long-histories of repeated, unsuccessful force-feeding; it is therefore understandable that they have little belief that further treatment will have any benefit. In the cases since *Re E*, clinical and judicial thinking appears to have been more appreciative of this line of reasoning, as shown by the rise of therapeutic jurisdiction. Secondly, that it is detrimental to the relationships they value. Ms X prioritised spending whatever time she had left with her grandfather; Z requested to be allowed to stay at home with her parents; and E felt like a burden to her family and wanted to protect their emotional wellbeing.[168] This is supported by research that found that what mattered most to those with anorexia was the nature of their relationship with their parents and health professionals.[169]

The parent's feelings are often expressed as those of helplessness, or a determination to advocate for their child. This was particularly poignant in *Re E* where E's parents made a statement describing their eighteen-year struggle to help their daughter. They movingly expressed their love and understanding of her torment, and advocated for her 'right to die', pleading that she could be granted "some control over what would be the last phase of her life".[170] The familial nature of anorexia has been described as a key feature of the condition. Giordano notes that experts regard it as a "systemic condition", which "involves the family in a profound and particular way".[171] She states that at the point where force-feeding is being considered, the family are morally entitled to be part of the decision and have their feelings considered.[172] The emphasis on relationships and emotion in the reasoning of the patient and their families accords with the moral theories promoted

[166] ibid [48], [51].
[167] *Re W (Medical Treatment: Anorexia)* [2016] EWCOP 13 [29], *Re E* (n 6) [5].
[168] *Ms X* (n 142) [5], *Cheshire v Z* (n 143) [14], *Re E* (n 6) [76].
[169] Jacinta Tan and others, 'Attitudes of patients with anorexia nervosa to compulsory treatment and coercion' (2010) 33 International Journal of Law and Psychiatry 13, 13.
[170] *Re E* (n 6) [80].
[171] Simona Giordano, 'Anorexia and Refusal of Life-Saving Treatment: The Moral Place of Competence, Suffering, and the Family' (2010) 17(2) PPP 143, 148.
[172] ibid.

by feminist theorists. Historically, courts acknowledged these relationships but did not give sufficient weight to them as core to moral reasoning.

The courts response has shown significant change since the controversial ruling in *Re E.* Peter Jackson J made frequent comments on E's intelligence and positive qualities, this seemed in stark contrast to E's evaluation of her own life as pointless.[173] Furthermore, Jackson J's decision to force-feed E was reasoned as a compassionate act to "protect her right to life under Article 2".[174] Yet, by repeatedly mentioning her intelligence he placed greater moral obligation on the court to save her life. This imposed meaning on her life that she, and her parents, did not recognise. This could also be viewed as an act of frustration, as her anorexia is perceived to be stifling her potential. Further, the feelings of the clinicians and her parents contradict. Though Dr Glover (the court-appointed expert) appreciated that E and her family were psychologically prepared for her death he still felt it to be in her best interests to be re-fed.[175] Cases since *Re E* have shown greater appreciation of the anorexic's feelings and reasoning. For example, King J went to great lengths to prioritise minimising Ms L's distress and understanding her views, and those of her family.[176]

The current trend towards therapeutic jurisdiction and away from force-feeding, demonstrates a change in legal reasoning. Though there may be other factors that can be attributed to this trend, the judgement in *Re E* gained considerable press attention and criticism.[177] The response has provoked courts to weigh factors differently, with the patient's emotional wellbeing at the heart of the decision. This demonstrates how cultural influences have impacted the reasoning and moral decision-making made in court. The degree to whether this influence can be said to be beneficial to the rationality and morality of the decision requires further exploration.

V. Are Emotions In Case Law Rational And Moral?

The case law discussed will now be considered in reference to the theories of emotion. To determine if they demonstrate that emotion can be rational and moral in these instances; I will consider whether the individuals in these cases can be said to be acting in accordance with moral emotions. If emotion is fundamental

[173] *Re E* (n 6) [5], [75], [132].
[174] ibid [141].
[175] ibid [87].
[176] *The NHS Trust v L* (n 141) [12].
[177] Daniel Sokol, '"As hard as it gets": the case of anorexic E and the right to die' The Guardian (London, 19 June 2012) <www.theguardian.com/law/2012/jun/19/anorexia-e-right-to-die> accessed 15 August 2017.

to our moral reasoning, as previously argued, then I shall ask whether emotion is sufficiently considered in case law and legislation. Before tackling these questions, let us return to the concepts of rationality and morality, and apply them to medical law. By applying these concepts, I will determine the role for emotion in cases of PDC and anorexia.

Legal enterprise sets out rules to guide human conduct. Rationality provides the means of creating these rules, and by which they are held accountable.[178] Brownsword takes a morally-driven approach to law, and has stated that for it to be rational it must meet two principles; firstly, it must be consistent, and secondly, it should be instrumental in guiding action.[179] These principles derive from traditional theories of rationality and moral judgement that advocate impartiality and reject a role for emotion.[180] As has been discussed, the role of emotion in rationality and morality is argued to be implicit to our moral reasoning. In comparing Brownsword's view to the account of moral emotions set out by Gibbard, one can see similarities in their principles despite differing accounts of the role of emotion. Gibbard describes certain 'moral emotions', such as guilt and anger, that produce societal moral norms.[181] These may not be fully engaged when one makes a decision, but a complete awareness of them is nonetheless essential to act morally.[182] The comparison with Brownsword arises from the Gibbard's claim that 'feelings' are normative, and ought to serve the agent as a guide to promote good. He also states that they ought to be "put to work" when accepted to be guides. He expands by saying that they must be harnessed to promote good.[183] As Gibbard has described, the requirement for consistency could be argued not as a reason to reject emotion in moral decision-making, but as a promotion of the appropriate use of moral emotions. Furthermore, emotions can therefore act as a means of guiding moral action in law.

A. Regarding prolonged disorders of consciousness

Kantian ethics have been very influential in philosophy and law, and have put forward a different approach to how one should be guided to act morally. Kant

[178] Richard Huxtable, Law, Ethics and Compromise at the Limits of Life: To Treat or Not to Treat? (Routledge 2013) 30.
[179] John Adams and Roger Brownsword, Key Issues in Contract (Butterworths 1995) 10.
[180] Huxtable (n 178) 26.
[181] Allan Gibbard, Wise Choices, Apt Feelings: A Theory of Normative Judgement (OUP 1990) 126–127.
[182] ibid.
[183] ibid 274.

would argue that a judge should only be motivated by duty. The judiciary have a responsibility to act in accordance to the law and morals. In this sense, their conduct follows a deontological ethic, as it is the permissibility of the act that is morally significant. For example, they have a duty to respect the sanctity of life and should not permit acts that contradict this principle. Kant would argue that their conduct should be in accordance with the categorical imperative. The principle formulation for which is as follows: "act only according to the maxim whereby you can at the same time will that it should become a universal law".[184] Therefore, to test whether a moral action is permissible, one must determine whether there would be a contradiction in conception if it were universalised. For example, it follows that it is wrong to make lying promises: if everyone were to make promises they would not keep, it would undermine the purpose of making a promise.[185] The historical requirement of judicial approval in withdrawal of clinically assisted nutrition and hydration in PDC cases can be viewed as an act that became standardised in law. For those in a vegetative state, the courts favoured withdrawal of life-sustaining treatment and thereby deemed this act to be morally permissible. This is due to the categorisation of withdrawal as an omission, rather than an act that breaches the sanctity of life principle. However, the failure here is that the strict adherence to deontological morality results in a disregard for the wider implications of this action. This includes consideration of the emotional implications, by which I mean the suffering of families who will still feel they have no humane means to end their loved one's suffering.

An alternative moral position to apply to PDC cases would be that of an ethic of caring. In applying Noddings' ethical theory, the position of a judge is one of 'caring-about' the patient. 'Caring-about' "involves a certain benign neglect", it is more public than personal, and does not call for the "engrossment, commitment, displacement of motivation" involved in caring for someone.[186] The 'one-caring' "involves a special regard for the 'cared-for', and "stepping out of one's own personal frame of reference into the other's".[187] The 'one-caring' has a much deeper sense of the patient's needs and point of view. The withdrawal of clinically assisted nutrition and hydration as the only means of death available neglects the 'one-caring'. This is shown by the views of families reported in the Kitzinger and Kitzinger study, and vast number of cases that were not taken to court prior to the recent legal developments. The court's reluctance to tackle the issue of the means

[184] Kant (n 16) 30.
[185] Driver, Ethics (n 11) 88.
[186] Noddings (n 47) 112.
[187] ibid 24.

of death demonstrates the disconnect between the emotional needs and reasoning of families, and the moral reasoning of the courts.[188] Reasoning in this instance requires a deep sense of compassion for those involved, the moral reasoning advocated by Kant cannot account for the deeply personal nature of these cases. Furthermore, the basis for moral permissibility attributed to withdrawal, rather than say lethal injection, has a deontological basis. Yet, on a caring basis the moral reasoning could be argued very differently. The emotional impact of watching the patient die from starvation on the 'one-caring' is evidently much greater than that of an instant death. The emotional repulsion may vary according to proximity; for example, a stranger may be horrified to hear of euthanasia, whereas those involved in the caring may view this as compassion. Nevertheless, a greater consideration of the emotions may give a different weight to the moral reasoning in these cases.

A utilitarian would argue that the court's moral reasoning is right when it will result in the overall greatest happiness. Therefore, the unhappiness of families who cannot withdraw clinically assisted nutrition and hydration, due to the means of death, should be considered. However, the practicality of considering the views of persons whose voices have not been heard in court, questions its pragmatic validity. A judge will naturally be more emotionally invested in the case before them, rather than in those hypothetical cases. Where the well-being of other persons is not under scrutiny in the trial. Furthermore, courts are under increasing pressure to attribute greater weight to the wishes and feelings of incapacitated patients, in line with the approach taken by the UNCRPD. This will result in greater opportunity for the judge to understand the emotions of the patient and their families. Placing the wishes of the patient at the heart of best interests decision-making may result in a move away from utilitarian moral ideals. Prioritisation of the patient's wishes forces the judge to prioritise the 'happiness' of the patient over others. This was demonstrated in *Wye Valley NHS Trust v B,* a landmark case in its approach to best interests decision-making.[189] Mr B was a 73-year-old gentleman who suffered from schizophrenia and had a severely infected leg that required amputation to save his life. Mr B was adamant that he did not want his leg amputated, but he lacked capacity to make this decision. Peter Jackson J made the rare decision to meet Mr B himself, to gain a better understanding of his best interests. He refused to grant the application that it was in Mr B's best interests to amputate his foot, against the advice of expert evidence. Here we can see that the morally right outcome was deemed to be that which protected Mr B's autonomy, rather than supporting medical advice. The moral reasoning here is not demonstrated by a utilitarian or

[188] Kitzinger and Kitzinger (n 96).
[189] *Wye Valley NHS Trust v B* [2015] EWCOP 60.

Kantian approach, but morality that is derived from caring for Mr B's emotional wellbeing.

B. Regarding anorexia

The rise of therapeutic jurisdiction observed in anorexia cases can also be considered as a result of a deeper understanding of the patient's emotional reasoning. It is questionable whether this would be justified by Kantian ethics: one would need to consider whether it is permissible to stop enforced feeding on patients with anorexia. There may be patients for whom force-feeding is an important part of their recovery, and who would deteriorate without that treatment. Perhaps in this instance the utilitarian approach is better appeased. In *Re E* the decision to force-feed went against the reasoning of E, her parents, and the clinicians who had worked with her. Their collective happiness may have been greater if the outcome had been to respect E's wishes. However, it could be argued that the rise in therapeutic jurisdiction in anorexia cases following *Re E* may have other negative outcomes. Where there may be greater reluctance in the courts and hospitals to force-feed patients who would benefit. This could result in poorer outcomes for patients who do not end up in court and the overall net happiness would decrease. However, it would not be possible to foresee these consequences if one were attempting to act following a utilitarian ethic. I would argue that the moral permissibility of therapeutic jurisdiction is best reasoned under an ethic of care and emotion. The revulsion of forced treatment that is perceived to be futile is a key driver in the court's reasoning. Reasoning that tries to act 'objectively' and does not give weight to the emotions of those involved, as it was not deemed to be in accordance with judicial impartiality, seems to be out of step with wider changes in medical law.

Prinz defends the position that emotions are essential in moral theory.[190] He posits that "moral emotions promote or detect conduct that violates or conforms to a moral rule".[191] Reactive moral emotions, such as indignation at injustice, are central to our morality.[192] The emotional response to force-feeding is a moral response to the agent's detection of a rights violation. The right to bodily integrity and autonomy is threatened and the patient's response is to act in a way to defend themselves. Reflexive moral emotions, such as guilt, occur when one has violated the autonomy rule of a person one cares about.[193] By meeting the patient, the judge acts to raise the importance of the patient's wishes and feelings, but it will

[190] Jesse Prinz, The Emotional Construction of Morals (OUP 2007) 13.
[191] ibid 68.
[192] ibid 70.
[193] ibid 76.

also increase the judge's caring attitude. In Noddings' terms; this moves the judge from a 'caring-about' position closer to 'caring-for'. This in turn will result in a greater emotional response if one acts in a way that violates moral rules. Guilt can develop into shame, which arises when "cultures start to label certain acts as unnatural or deviant".[194] Societal changes that expect greater weight to be given to the patient's autonomy and aversion to paternal medical practices. The act of force-feeding has become increasingly more repulsive and morally impermissible.

The rise of therapeutic jurisdiction in anorexia cases demonstrates a change in how moral judgments are formulated. I have argued that our emotional response is core to our moral reasoning, and that societal emotional reactions guide our morality. This is best explained by an emotional and caring basis to human morality. The outcome of PDC trials also demonstrates how repulsion to starvation, the antithesis of the caring attitude, results in a preference for withdrawal in vegetative state but not minimally conscious state. The likelihood of neurological recovery in minimally conscious state cases is minimal. Yet the idea of starving a person with even a minimal level of consciousness provokes greater emotional repulsion, and I have argued that this is what underpins decisions to continue feeding.

VI. Conclusion

Utilitarian and Kantian rationalists will argue that moral judgements are made through reasoned decision-making. An impartial decision-maker considers all the information, then weighs and tests it according to moral principles. Nonetheless, what Kant and Singer failed to account for is how emotions are intrinsic to moral decision-making. Emotions not only help us to focus our attention on important matters, they help us to process information quickly, and are active in how we formulate our judgement. Moral decision-making is based on rules that we have learnt through our emotional experiences. Our emotions are a prerequisite to reasoning and provide a means of evaluating whether we have acted morally. When we feel injustice, we must learn how this should inform moral action to restore justice. Those who cared for Tony Bland felt outraged by his circumstances. They acted on his behalf to rectify the injustice that had befallen him. Reason without emotion cannot account for this. Reason alone might rationalise that we must preserve the sanctity of life and he was not actively suffering therefore should continue living, conversely reason could argue that he had no meaningful life therefore one should withdraw. However, it is our emotional evaluation of these

[194] ibid 78.

options that inform how we judge them. It is from this perspective that I have justified an internal role for emotion.

Cultural influences on how we interact with reasoning is also important to consider in these cases. These may change over time, and our emotions are responsive to these changes. For example, the Law Commission has recommended that greater weight be given to the patient's wishes and feelings in accordance with the Human Rights Act 1998 and UNCRPD.[195] Recent legal developments regarding the determination of best interests in both PDC and anorexia cases demonstrate a change in judicial reasoning to this effect. By giving greater weight to the patient's wishes and feelings the decision-maker is encouraged to show a greater understanding the emotional life of the individual. The turn to therapeutic jurisdiction in anorexia cases is not just a sign of a cultural shift towards autonomy, but also towards emotion. Although, the influence of emotion in PDC cases seems to be at odds with societal opinion. Families have reported they would support euthanasia, but not withdrawal of clinically assisted nutrition and hydration. The means of death provokes different emotions, and the moral reasoning in each eventuality is very different. The emotional response to murder is quite different to that of compassionate euthanasia. Perhaps if legal reasoning in these cases had been more informed by emotion, there would be greater agreement with societal and cultural influences.

Institutions impose their own specific arrangements that influence emotional reasoning. Institutions, such as courts and hospitals, differ in how they "model, direct, and constrain the psychological/emotional repertoire".[196] Some institutions may valorise some emotions, and stigmatise others. The cultural institutional differences require specific consideration in the feeding decisions discussed. The emotional response, reasoning, and morality, within a healthcare setting will be different to that of a court. Hospitals promote moral action based on caring and beneficence. Courts seek moral behaviour based on justice and evidence. Individuals entering these domains will in turn be influenced by the differing perspectives. *Re E* demonstrated a cultural difference in the emotional reasoning of E's clinicians and the court. These differences need to be recognised and analysed further, as this influences how moral decisions are reasoned and whether they are right.

In this article, I have defended emotions as assuming moral importance. Medical law concerning feeding decisions at the end-of-life have been described and examined. In applying the concepts of the philosophy of emotion to these legal

[195] Law Commission, Mental Capacity and Deprivation of Liberty (Law Com No. 372, 2017) para 14.12.

[196] Amélie Oksenberg Rorty, 'Enough already with "Theories of the Emotions"' in Robert Solomon (ed), *Thinking about Feeling: Contemporary Philosophers on Emotion* (OUP 2004) 277.

issues I have responded to my research question: Should emotion be a component of how we make decisions about feeding at the end-of-life? A better understanding of how emotions support our rationality and morality will help us to make better decisions. End-of-life feeding decisions are particularly emotive, and are at greater risk of being reasoned from a traditional rationalist perspective. The person's emotional response will vary according to the role and association to the case. Yet, without understanding the emotional forces in one's reasoning and morality, one cannot truly make a fully informed decision. This should inform medical law.

Access to Legal Advice: Should PACE Go Further or Take a Step Back?

Clinston Chiok*

I. Introduction

At present, the Police and Criminal Evidence Act 1984 (PACE) recognises the right of suspects to be given access to legal advice before being interviewed by the police and permits solicitors to be present during the police interview. This article analyses whether, in seeking to comply with Article 6 of the European Convention of Human Rights (ECHR), the current approach of PACE goes too far, or ought to be strengthened further.

The first part of the article will address whether Article 6 includes a right to legal advice prior to police interview and, if so, whether PACE ought to strengthen this right by making it mandatory for suspects to speak to a solicitor before they are interviewed by the police. The second part will address whether PACE goes beyond the United Kingdom's (UK) Article 6 obligations by permitting solicitors to be present during the police interview. This part will also address practical concerns about whether the presence of solicitors will encourage 'no comment' interviews, which are said to impede police investigation and ultimately be against the interests of the defence.

II. Right To Legal Advice Prior To Police Interview

A. Is the right to legal advice prior to interview required under article 6?

Under the ECHR, Article 6 paragraph (3)(c) provides for the right to legal assistance, which is fundamental to its guarantee of the right to a fair trial under Article 6.[1] Enshrining this under domestic law, section 58(1) of PACE provides for

* LL.B. (Nott); LL.M. (Cantab). I am grateful to the editors for their helpful comments. Any errors remain my own.

1 *Salduz v Turkey* (2009) 49 EHRR 19 [50].

the right of a suspect arrested and held in custody to "consult a solicitor privately at any time". Subject to exceptions, the suspect will generally not be interviewed until such advice has been received.[2] However, does PACE adopt a wider scope than what Article 6 of the ECHR requires?

First, one may argue that the right to a fair trial under Article 6 only applies to a criminal charge and not during the preliminary investigatory phase. However, both domestic courts and the European Court of Human Rights (ECtHR) have recognised that the term 'criminal charge' is an autonomous concept, satisfied where a suspect is "substantially affected" by steps taken against him.[3] In *John Murray v United Kingdom*, the ECtHR found Article 6 to apply during the "preliminary investigation into an offence by the police", since Northern Irish legislation allowed adverse inferences to be drawn at trial from the suspect's silence during interrogation.[4]

Since England and Wales' Criminal Justice and Public Order Act 1994 (CJPOA) is similar to the Northern Irish legislation which was in consideration in *Murray* in allowing adverse inferences to be drawn, the English court in *R v Stratford Justices, ex p Imbert* accepted that Article 6 would apply during the preliminary investigation stage.[5] *Murray*[6] was re-emphasised in *Salduz*, where the ECtHR expressed that "as a rule, access to a lawyer should be provided as from the first interrogation of a suspect by the police", deeming such access to be fundamental to the right to a fair trial under Article 6.[7]

Even if Article 6 requires that legal access should be provided from the first interrogation of the suspect, does this require PACE to permit access to legal advice *before* any police interview has taken place? The concurring opinions in *Salduz* state that legal assistance ought to be provided "at the very beginning of police custody or pre-trial detention… not only while [the suspect is] being questioned".[8] This general principle was affirmed in *Dayanan v Turkey*.[9] But why is this so? In order to understand why such early access is fundamental to the right to fair trial under

[2] Home Office, 'Police and Criminal Evidence Act 1984 (PACE) – Code C: Revised Code of Practice for the Detention, Treatment and Questioning of Persons by Police Officers' (February 2017) [6.6] <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/592547/pace-code-c-2017.pdf> accessed 13 May 2017.

[3] *Deweer v Belgium* (1980) 2 EHRR 439 [46].

[4] *John Murray v United Kingdom* (1996) 22 EHRR 29 [62], [66].

[5] *R v Stratford Justices, ex p Imbert* [1999] 2 Cr App R 276, 285(g) (Buxton LJ).

[6] *Murray* (n 4).

[7] *Salduz* (n 1) [50], [55].

[8] ibid (concurring opinion of Judge Zagrebelsky, joined by Judges Casadevall and Türmen, concurring opinion of Judge Bratza).

[9] *Dayanan v Turkey App no 7377/03* (ECHR, 13 October 2009) [32].

Article 6, it is first necessary to consider the justification for the general existence of the right under this provision.

First, there can be more than one justification for the existence of Article 6.[10] In the decision of the UK Supreme Court in *Cadder v HM Advocate*, Lord Hope recognised that the right is a vital safeguard against ill-treatment whilst also accepting that it is closely linked with the protection against self-incrimination.[11] It will be argued that both the ECtHR and PACE's recognition of the right to legal advice prior to interview are based on the overarching notion of ensuring a fair procedure under Article 6.

In *Salduz*, the ECtHR stated that a suspect would be in a "particularly vulnerable position" during the interrogation stage, exacerbated by complex legislation on criminal procedure such as "the rules governing the gathering and use of evidence".[12] Such vulnerability could only be compensated by "the assistance of a lawyer whose task it is, among other things, to help ensure respect of the right of an accused not to incriminate himself".[13] This protection against self-incrimination was expressed in *Saunders v UK* to be a fundamental aspect of the notion of a fair procedure.[14]

Looking closer to home, the Royal Commission on Criminal Procedure 1981 (RCCP 1981) expressed a similar view. Prior to the enactment of PACE, a suspect's right to a solicitor was derived from the Judges' Rules.[15] It was subject to exceptions, rarely exercised, and most suspects did not know such a right existed.[16] Even if they did, there was limited legal aid available for police station legal advice.[17] The right to legal advice was identified by RCCP 1981 as a key safeguard that could have prevented miscarriages of justice such as the Confait affair,[18] where there was a wrong conviction for murder on the basis of confessions made by mentally impaired teenagers under oppressive police questioning. Enshrining the right to legal advice was thought to protect vulnerable suspects against self-incrimination and more broadly, to balance "the interests of the community in bringing offenders

[10] John Jackson, 'Responses to Salduz: Procedural Tradition, Change and the Need for Effective Defence' (2016) 79(6) MLR 987, 1001.

[11] *Cadder v HM Advocate* [2010] UKSC 43, [2010] 1 WLR 2601 [33], [44].

[12] *Salduz* (n 1) [54].

[13] ibid.

[14] *Saunders v UK* (1997) 23 EHRR 313 [68].

[15] Ed Cape, 'The Rise (and Fall) of a Criminal Defence Profession' [2004] Crim LR 72, 81.

[16] Henry Fisher, 'Report of an Inquiry into the Circumstances Leading to the Trial of Three Persons on Charges Arising out of the Death of Maxwell Confait and the Fire at 27 Doggett Road, London SE6' (London: HMSO, 1977).

[17] Cape (n 15) 82.

[18] *R v Lattimore* (Colin George) (1976) 62 Cr App R 53.

to justice and the rights and liberties of persons accused or suspected of crime",[19] ensuring a fairer procedure prior to trial. Thus, both the ECtHR and PACE's justification for the provision of legal advice *during* the interrogation stage is to protect the suspect from self-incrimination, which falls under the overarching notion of a fair procedure.[20]

Since the justification for legal advice is identified as ensuring a fair procedure, does it warrant the right to legal advice *prior* to interview? Fundamental to the notion of a fair procedure is the privilege against self-incrimination and right to silence.[21] In *Cadder*, Lord Rodger pointed out that the right to legal advice before being interviewed was "derived from the need for legal assistance for other purposes", and not for protection against self-incrimination.[22] It is submitted that one of the "other purposes" that Lord Rodger alludes to is the protection of the right to silence.

Leverick identifies that protecting the right to silence involves the legal adviser fulfilling two distinct duties prior to the interview: (i) ensuring that the suspect comprehends the right to silence, and (ii) aiding him in determining his best interests.[23] Duty (i) entails explaining the nature of the right to silence and how adverse inferences may work against him, while duty (ii) relates to assisting suspects "to make an informed choice about their best interests on the basis of the [right to silence] and the nature of the evidence against them".[24]

Consistent with the two duties identified by Leverick, the ECtHR in *Dayanan* expressed the view that "the fairness of proceedings requires that an accused be able to obtain the whole range of services specifically associated with legal assistance", including "discussion of the case" and "preparation for questioning".[25] Lord Rodger also noted the link between protecting the right to silence and provision of legal advice prior to interview in *Cadder*, where he stated that such legal advice assisted in deciding "whether [the suspect] should say anything at all and, if so, how far he should go".[26] As explained above, the provision of legal advice is justified on the basis of ensuring a fair procedure.

Therefore, it follows that the ECtHR and PACE are correct to recognise the right to legal advice before being interviewed by the police: it protects the suspect's

[19] Royal Commission on Criminal Procedure, The Investigation and Prosecution of Criminal Offences in England and Wales: The Law and Procedure (Cmnd 8092, 1981).

[20] *Saunders* (n 14) [68].

[21] ibid.

[22] *Cadder* (n 11) [70].

[23] Fiona Leverick, 'The Right to Legal Assistance During Detention' (2011) 15(3) Edinburgh Law Review 352, 366.

[24] ibid 369.

[25] *Dayanan* (n 9) [32].

[26] *Cadder* (n 11) [92].

right to silence, which is fundamental to the notion of a fair procedure under Article 6 of the ECHR.[27]

B. Should access to legal advice at the pre-interview stage be mandatory?

Given the importance of legal advice prior to interview, should PACE be strengthened by making it mandatory for the suspect to speak to a solicitor before an interview?

As Easton notes, there is no nation-wide data on request rates for legal advice.[28] Bucke and Brown's study in 1997 indicates a 40% request rate,[29] while Pleasence and others' study in 2011 suggests that the rate has increased to 44.9%.[30] Although Skinns' research in 2009 states that request rates for legal advice were about 60%,[31] Pleasance and others assert that this is not indicative of general request rates, since Skinns' study was based on only two police stations and there was a large variation (17%) on request rates between them.[32] However, Skinns' results corroborate studies done by Pleasance and others[33] and Brown[34] in illustrating that rates of request for legal advice vary considerably between police stations.

Brown suggests that the variation in request rates between stations may be due to arrangements and practices between them or due to differences in interpretation of PACE and the relevant Code of Practice provisions.[35] Pleasence and others note that a crucial driver of requests for legal advice is the seriousness of the offence,[36] which corroborates Maguire's early findings that detainees for minor offences do not generally request legal advice, since it was not seen as "likely to be of any benefit".[37] Additionally, in the time between Brown's 1997 study and Pleasence

[27] *Saunders* (n 14) [68].

[28] Susan Easton, Silence and Confessions: The Suspect as the Source of Evidence (1st edn, Palgrave Macmillan 2014) 93.

[29] Tom Bucke and David Brown, In Police Custody: Police Powers and Suspects' Rights under the Revised Codes of Practice (Home Office Research Study No. 174, 1997).

[30] Pascoe Pleasence, Vicky Kemp and Nigel Balmer, 'The Justice Lottery? Police Station Advice 25 Years on from PACE' [2011] Crim LR 3, 10.

[31] Layla Skinns, 'I'm a Detainee: Get Me Out of Here' (2009) 49(3) British Journal of Criminology 399.

[32] Pleasence and others (n 30) 12.

[33] ibid.

[34] David Brown, Detention at the Police Station under the Police and Criminal Evidence Act 1984 (Home Office Research Study 104, 1989).

[35] ibid 73.

[36] Pleasence and others (n 30) 11.

[37] Mike Maguire, 'Effects of the PACE Provisions on Detention and Questioning' (1988) 28(1) British Journal of Criminology 19, 31.

and others' 2011 findings, it is startling that the percentage of those who requested and actually consulted with a solicitor grew only from 34%[38] to 36.5%.[39] Given that the right to legal advice before the interview protects the suspect's right to silence and gives effect to a fair procedure, legal advice must not be precluded due to "arbitrary factors" such as the police station that the suspect finds himself in.[40] To prevent such a possibility from occurring and to strengthen the right to legal advice, there is a compelling argument that legal advice should be made mandatory before suspects are engaged in an interview.

However, it may be counter-argued that in reality, mandatory legal advice prior to the interview may not significantly strengthen the right to legal advice under Article 6. As stated above, one of the purposes of legal advice is to assist the suspect in identifying his best interests. In the context of deciding if the right to legal advice should include the presence of the solicitor in the interview room in order to identify the suspect's best interests, the minority of the Canadian Supreme Court in *R v Sinclair* pointed out that a solicitor is unlikely to be able to give effective advice on the suspect's best interests during the pre-interview stage, as the solicitor would have insufficient information about the evidence against the suspect.[41] Moreover, the suspect's best interests may only become apparent as questioning develops.[42] Therefore, mandatory legal advice may not significantly strengthen the right to advice. However, it would at least assist the suspect in understanding his right to silence, fulfilling the other objective of the right to silence at the pre-interview stage.

II. Right to Legal Advice During Police Interview

A. Does article 6 require PACE's present approach?

Does PACE go too far in permitting solicitors to be present in the police interview? To address this, the justification behind the right to legal advice must first be analysed.

As Jackson points out, the Canadian Supreme Court in *Sinclair*[43] decided that the right to legal advice is merely "informational rather than protective".[44] Therefore, a solicitor's presence during the interview was unnecessary as its role could be discharged prior to the interview. In contrast, since PACE's justification for legal advice is to ensure a fair procedure (as argued above), it is submitted

[38] Brown (n 34) 8.
[39] Pleasence and others (n 30) 10.
[40] ibid 6.
[41] *R v Sinclair* [2010] 2 SCR 310 [104] (Binnie J).
[42] ibid [87] (Binnie J).
[43] ibid.
[44] Jackson (n 10) 1006.

that presence of a solicitor during the police interview is necessary. In *Salduz*, the ECtHR held that "as a rule, access to a lawyer should be provided as from the first interrogation of a suspect by the police".[45] Lord Hope in *Cadder* explained that the rule in *Salduz* was laid out "to protect a detainee against duress or pressure of any kind that might lead him to incriminate himself" and "to help to ensure that the right of an accused not to incriminate himself is respected".[46] This necessitates a lawyer's presence and advice during the interrogation in order to protect the suspect's privilege against self-incrimination, thus ensuring a fair procedure.

Despite permitting a solicitor's presence during the interview, in substance, it is submitted that PACE, in fact, falls short of ensuring that such presence secures a fair procedure. As stated above, the right to a fair trial under Article 6 applies at the interrogation stage. Fundamental to Article 6 is the equality of arms,[47] which seeks to ensure that "procedural resources enjoyed by the parties are fairly matched" and "must be respected at each stage… where Article 6 is found to be applicable".[48] It is submitted that this is not met due to lack of evidential disclosure and effective legal assistance.

The first argument is that the interrogation stage lacks procedural clarity in relation to evidential disclosure, infringing the notion of equality of arms. First, the primary legislation on disclosure, the Criminal Procedure and Investigations Act 1996 (CPIA), does not govern disclosure by the police and leaves a lacuna in terms of statutory coverage. In addition, PACE and Code C "offer no guidance to the police or defence lawyers on the use of evidence at custodial interrogation".[49] Although the courts have implied that the extent of disclosure required depends on the factual complexity of the case, it has avoided elaborating on how this is to be decided.[50] Without knowledge of the available evidence against the suspect, it is unlikely that the solicitor present during the interrogation can provide any effective advice, rendering the right merely theoretical and illusory.

Secondly, it may be argued whether PACE complies with Article 6 paragraph (3)(c) when providing for legal assistance during the police interview. Jackson observes that more cases are being disposed of before reaching trial; the trial has effectively been brought forward to the investigatory phase, since "formal sentence after trial is being replaced by a negotiated sentence or sanction after investigation".[51] Furthermore, the result of CJPOA and the new caution is that

[45] *Salduz* (n 1) [55].

[46] *Cadder* (n 11) [33].

[47] *Jasper v UK App no 27052/95* (16 February 2000) [51].

[48] Raymond Toney, 'Disclosure of evidence and legal assistance at custodial interrogation: what does the European Convention on Human Rights require?' (2001) 5(1) International Journal of Evidence & Proof 39, 48.

[49] ibid 52.

[50] *R v Argent* [1997] 2 Cr App R 27.

[51] Jackson (n 10) 1017.

suspects are accountable at a much earlier stage than at trial.[52] It follows that relevant procedural safeguards, such as qualified legal assistance, ought to be brought forward to the investigatory phase. However, for purposes of PACE, a solicitor is defined wider than in the traditional sense, including "an accredited or probationary representative" regulated by the Legal Aid Agency.[53] Would such persons meet the requirement for the provision of legal assistance in Article 6(3)(c)? Although Harris[54] argues that Article 6(3)(c) does not necessitate the use of qualified lawyers, due to the Convention's legislative background and the use of the term "legal assistance" instead of "qualified representative", it is noted that legal assistance must still rise to the level of "effective assistance".[55] Given that the trial is effectively brought forward and taking into account the complexities introduced into custodial interrogation by CJPOA, it is submitted that "only an experienced and robust defence at custodial interrogation will ensure that the suspect's best interests are fully represented"[56] and that PACE's definition of a solicitor may not satisfy Article 6(3)(c) and thus may not provide for equality of arms.

Therefore, it is submitted that PACE does not go too far in allowing a solicitor's presence during the interview, because it is necessary to ensure a fair procedure. Instead, it is arguable that PACE falls short in ensuring that the substance of the right is effective.

B. Does a solicitor's presence encourage 'no comment' interviews?

Historically speaking, Moston[57] notes that 'no comment' interviews are more likely for suspects who have legal advice than for those who do not. Given Ashworth's observations that the number of suspects "having an adviser with them, has risen dramatically",[58] and given that Pleasence and others' recent research indicates that request rates for legal advice have risen to 44.9%,[59] one would expect that the number of no comment interviews ought to have increased. However,

[52] John Jackson, 'Silence and Proof: Extending the Boundaries of Criminal Proceedings in the UK' (2001) 5 International Journal of Evidence & Proof 145, 168.

[53] Code C (n 2) [6.12].

[54] David John Harris, Michael O'Boyle and Colin Warbrick, Law of the European Convention on Human Rights (1st edn, Butterworths 1995) 265.

[55] *Artico v Italy* (1981) 3 EHRR 1 [34].

[56] Toney (n 48) 40.

[57] Stephen Moston, Geoffrey Stephenson and Thomas Williamson, 'The Effects of Case Characteristics on Suspect Behaviour During Questioning' (1992) 32(1) British Journal of Criminology 23, 36.

[58] Andrew Ashworth and Mike Redmayne, The Criminal Process (4th edn, OUP 2010) 244.

[59] Pleasence and others (n 30) 10.

whilst comparing almost the same police stations, Bucke notes that the percentage of no comment interviews decreased from 10% in 1998 to 6% in 2000.[60] In addition, Bucke found that one of the areas which has seen the greatest decline in the use of silence was from suspects who received legal advice; and amongst that group, the proportion who engaged in no comment interviews dropped from 20% to 13%.[61] Thus, recent studies do not support the view that a solicitor's presence would encourage no comment interviews.

The decline in no comment interviews where legal advice is sought can be explained by two closely related reasons. First, the enactment of CJPOA means that under the new caution, adverse inferences can be drawn from a suspect's failure to mention facts to the police that he later relies on in defence.[62] Bucke observes that post-CJPOA, it is more probable that legal advisers will advise suspects to respond to questions.[63] As Jackson notes, keeping absolutely silent "could be a risky strategy as the case law is by no means clear as to how a court or jury should view a lack of disclosure".[64] Secondly, the enactment of CJPOA has altered the solicitor's previous practice of leveraging on the threat of silence in order to obtain information from the police.[65] Post-CJPOA, executing the threat of silence may harm the defence's interests through adverse inferences. As Sanders explains, "neither limited disclosure of the police case nor legal advice to remain silent necessarily insulates the suspect from adverse inferences".[66] Given the threat of adverse inferences and that the police are under no clear duty of disclosure (as explained above), Ashworth[67] notes that legal advisers now advise cooperation with the police; to get information from the police, the suspect needs to offer information in return.

However, in certain situations the solicitor may indeed encourage no comment interviews, such as when the case is "conducted by the police on the basis of fishing expeditions and there [is] no real evidence against the client".[68] Additionally, no comment interviews may not be significantly influenced by legal advice at all. Baldwin asserts that suspects are unlikely to change their initial position of silence

[60] Tom Bucke, Robert Street and David Brown, The Right of Silence: The Impact of the Criminal Justice and Public Order Act 1994 (Home Office Research Study No. 199, 2000) 31.
[61] ibid 32.
[62] Criminal Justice and Public Order Act 1994, s 34.
[63] Bucke and others (n 60) 25.
[64] Jackson (n 52) 160.
[65] ibid 169; David Dixon, 'Common Sense, Legal Advice and the Right of Silence' [1991] PL 233.
[66] Andrew Sanders, Richard Young and Mandy Burton, Criminal Justice (4th edn, OUP 2010) 268.
[67] Ashworth and Redmayne (n 58) 245.
[68] Jackson (n 52) 161.

or co-operation, despite the conduct of the interview.[69] This corroborates Bucke's[70] findings that professional criminals and terrorists are still unlikely to answer police questions despite the provision of legal advice and the introduction of CJPOA. Therefore, it is submitted that generally, a solicitor's presence discourages no comment interviews, but in certain situations a solicitor may encourage silence or have no effect on the suspect.

C. Do 'no comment' interviews impede police investigations?

In considering whether the right to silence ought to be modified, the Runciman Commission noted the police's view that the suspect's refusal to answer questions would "seriously impede the efforts of investigators to fulfil their function of establishing the facts of the case".[71] By speaking during the interrogation, it was asserted by the police that it allowed them to seek the truth by confirming any exonerating assertions by the suspect and to "direct [their] attention towards the guilty".[72] A refusal to speak prevents any such evidence from emerging. This concurs with Bentham's sentiment more than a century ago, that the suspect is "the most satisfactory species of evidence" since he would not speak falsely to his own detriment.[73] Similarly, Bentham argued that as much "light of evidence" should be let in since "the end it leads to, is the direct end of justice, rectitude of decision".[74]

Despite the above, it is submitted that no comment interviews do not, in fact, impede police investigation. As Jackson states, "just as the right to silence can be grossly exaggerated as a mechanism for protecting the innocent, it can also be grossly exaggerated as an obstacle for convicting the guilty".[75] The police's argument is premised on the fact that the suspect is the best source of evidence in the police's supposed search for truth, which may not necessarily be true. First, given the availability of other forms of evidence such as video footage or forensic evidence, Easton argues that there are few situations where the suspect's account is the sole or best source of evidence.[76] Next, suspects may not be the best source of evidence since the pressures of interrogation may taint the quality of the evidence.[77] This is backed up by Bucke's findings, where some police officers commented that

[69] John Baldwin, 'Police Interview Techniques: establishing Truth or Proof?' (1993) 33 British Journal of Criminology 325, 333.

[70] Bucke and others (n 60) 36–37.

[71] Royal Commission on Criminal Justice, Report, Cm 2263 (1993) 51.

[72] ibid.

[73] Jeremy Bentham, The Works of Jeremy Bentham: Published under the Superintendence of His Executor, John Bowring (W Tait; Simpkin, Marshall & Co 1843) 451.

[74] ibid 336.

[75] John Jackson, 'Reconceptualising the Right of Silence as an Effective Fair Trial Standard', (2009) 59 International & Comparative Law Quarterly 835, 850.

[76] Easton (n 28) 14.

[77] ibid.

CJPOA's effect was to substitute silence for "a pack of lies" and that suspects "[lied] a little better… instead of saying no comment".[78] Furthermore, if no comment interviews impeded police investigations, one would expect an inverse relationship between the use of silence and admission rates or convictions secured. Instead, Bucke found that the decrease in use of silence has not affected the admission rates and convictions secured.[79] On the contrary, Williamson notes that it is more likely for suspects who are silent to plead guilty and that a majority of such suspects end up getting convicted.[80] This supports Leng's study, which indicates that only a minority of non-prosecuted cases and acquittals involved the exercise of silence, and those outcomes were not truly a product of silence itself.[81] Therefore, it is submitted that no comment interviews cannot be said to impede police investigation, since it is backed up by neither qualitative nor quantitative evidence.

D. Do 'no comment' interviews turn out to be against the interests of the defence?

As stated earlier, section 34 CJPOA allows adverse inferences to be drawn from the suspect's failure to mention facts when questioned, which he later relies on in his defence. Although the provisions state that a suspect shall not have a case to answer or be convicted solely on an inference,[82] the court in *Murray*[83] emphasised that a suspect cannot be convicted solely *or mainly* from such an inference. Thus, the statutory framework suggests that no comment interviews can be against the interests of the defence due to the operation of CJPOA, but it will "play no more than a supporting role" for the prosecution.[84]

How does the theoretical framework of CJPOA apply in reality? Firstly, Jackson notes that some prosecutors perceive the legislation as merely assisting them indirectly by weakening the defence case, despite the potential for such silence to form part of their case.[85] Further, prosecutors may not even rely upon the legislation. Such abstinence can be explained by several reasons; "tactical (the danger of over-kill), presentational (the risk of diverting the jury from the thrust of the case) and personal (some indicating that they believed the provisions were

78 Bucke and others (n 60) 32.

79 Bucke and others (n 60) 34, 66–67.

80 Thomas Williamson and Stephen Moston, 'The Extent of Silence in Police Interviews' in Greer and Morgan (eds), The right to silence debate: Proceedings of a Conference held at the University of Bristol on 27 March 1990 (Bristol Centre for Criminal Justice 1990) 36–43.

81 Roger Leng, The Right to Silence in Police Interrogation (Royal Commission on Criminal Justice Research Study No. 10, 1993).

82 Criminal Justice and Public Order Act 1994, s 38(3).

83 *Murray* (n 4).

84 Ian Dennis, 'Silence in the Police Station: The Marginalization of Section 34' [2002] Crim LR 25, 29.

85 Jackson (n 52) 162.

unfair)".[86] Despite the above, the prosecution believes that silence could "tip the balance" in their favour in borderline cases.[87]

Secondly, it is unclear what a jury makes of the provisions. Although there is a view that juries may "place too much weight" on adverse inferences,[88] Jackson asserts that the primary view is that juries do not use the provisions to "plug a large gap in the prosecution case", but merely to draw adverse inferences against the defence.[89] In contrast, practitioners perceived that the provisions had an insignificant impact on juries[90] and that they were hesitant to draw inferences where a defendant was silent during interrogation but testified in court, since the defendants were not silent in front of them[91]. Additionally, practitioners identified that it was difficult to get the jury to pay attention to the judge's directions.[92]

In relation to the judiciary, Bucke[93] identifies that judges in the Crown Court are divided in opinion between those who are receptive and those who are concerned with the provisions. Although there is a general consensus that a more consistent judicial attitude will result after a passage of time, practitioners have expressed concern that the present divergence in judicial attitude may result in an inconsistent application of the provisions.[94] Additionally, some of the senior judiciary appear to be resistant to the provisions. Lord Bingham commented in *Argent* that "Parliament in its wisdom has seen fit to enact [s 34]",[95] and subsequently expressed in *Bowden* that the provision "should not be construed more widely than the statutory language requires".[96] Dennis[97] notes that post-*Bowden* cases have indeed reflected a restrictive approach towards the provisions.

Therefore, considering the view of the jury, prosecution, and judiciary towards the provisions, it is submitted that no comment interviews can be against the interests of the defence, albeit to a limited extent. Given the limited effect of the provisions, Dixon aptly states that CJPOA was enacted "for more symbolic than instrumental reasons in order to regain political ground lost by safeguards provided under PACE".[98]

[86] ibid 153.
[87] Bucke and others (n 60) 45.
[88] Jackson (n 52) 157.
[89] ibid 157.
[90] ibid 155.
[91] ibid 156.
[92] ibid 157.
[93] Bucke and others (n 60) 60.
[94] ibid 60.
[95] *Argent* (n 50) 32.
[96] *R v Bowden* [1992] 2 Cr App R 176, 181.
[97] Dennis (n 84) 30.
[98] David Dixon, 'Politics, Research and Symbolism in Criminal Justice: The Right of Silence and the Police and Criminal Evidence Act' (1991) 20 Anglo-American Law Review 27.

III. Conclusion

This article has argued that PACE does not go too far in recognising the right to pre-interview legal advice, and that making such advice mandatory would strengthen the right to a certain extent. Further, PACE is justified in allowing the presence of a solicitor during the police interview and it does not encourage no comment interviews; even if it does, it does not impede police investigation and has a limited impact against the interests of the defence. However, there is a possible argument that PACE in fact needs to go further in order to comply with Article 6 during the interview stage.

The Right to Translation and Interpretation in Criminal Proceedings: Providing a Common Code Between the Defendant and the Court

Chara Chioni-Chotouman*

I. Introduction

Noam Chomsky defined linguistic competence as an inherent capacity of native speakers.[1] The native speaker, through his linguistic competence, can be creative and easily produce new sentences in a way that allows the listener to understand them. Non-native speakers do not have this linguistic creativity and cannot fully understand the multitude of meanings expressed by a native speaker. The consequent inequality between native and non-native speaker defendants necessitates the presence of a linguistic mediator in the context of criminal proceedings, who facilitates communication between the defendant and the court.

This article examines the right to translation and interpretation in criminal proceedings at the European level. It is argued that the participatory and communicative character of this right defines both its scope and the conditions for its exercise. It is also stated that the right to translation and interpretation is more than a mere procedural right, playing an important role in the administration of justice. At the level of the European Union (EU), the European Convention of Human Rights (ECHR)—and specifically Article 6 paragraph 3(e)—is extensively analysed in light of the European Court of Human Rights (ECtHR)'s case law. Issues regarding the translation of documents and the quality of the services provided, as well as the exclusion of the defendant from the subsequent costs and the waiver of the right, are also addressed. Moreover, after specifying the reasons

* Ph.D. Candidate in Criminal Law and Criminal Procedure (National and Kapodistrian University of Athens), LL.M. in Criminal Law and Criminal Procedure (National and Kapodistrian University of Athens).

1 Noam Chomsky, Language and Mind (3rd edn, CUP 2006) 23.

behind the need for uniform regulation within Member States, I analyse the Directive 2010/64/EU in detail. Special attention is paid to the interpretation of attorney-client communication and to the autonomous character of the right to translation of documents. The issue of the non-inclusion of a list of essential documents and that of effective quality control are addressed. Furthermore, I allege that the relevant case law of the Court of Justice of the European Union (CJEU) has revealed the inefficiencies and ambiguity of the Directive, primarily because of the vague terminology and the wide margin of discretion left to Member States. Lastly, the ECHR and the Directive are compared and contrasted.

II. The Nature of the Right to Translation and Intrerpetation: a Communicative Procedural Right

The right of translation and interpretation is primarily linked to the right to be heard. The latter constitutes a fundamental right that derives from the acknowledgment that the defendant is a rational and responsible agent.[2] In the absence of this right, a defendant would not be considered a participant in the criminal procedure.[3] The physical presence of the defendant is not sufficient in itself; it is essential that his active and effective participation in the procedure is guaranteed. The right to translation and interpretation is inextricably linked to the right of the defendant to follow the proceeding. In this case, there is a lack of a common code between the sender and the receiver making it impossible for any message to be decoded. Thus, the defendant cannot fully understand the procedure nor express his views. If this right is not granted, the trial cannot be 'fair'.

The connection of the right to translation and interpretation with the fairness of proceedings is particularly evident in its interrelationship with other rights. The right to translation and interpretation is a requirement for the implementation of other procedural rights, such as the right of confrontation and the right against self-incrimination. A defendant who is not familiar with the language of the proceedings is more likely to answer in a self-incriminating manner.[4]

The right to translation and interpretation serves not only to enforce the right to a defense, but also acts as an essential condition for the proper administration of justice.[5] The presence of an interpreter serves all parties, but most of all the court itself, which cannot communicate with defendants who speak a different

[2] Sarah J Summers, Fair Trials (Hart Publishing 2007) 19.

[3] *Stanford v UK* App no 16757/90 (ECHR, 23 February 1994), para 26; *Roos v Sweden* App no 19598/92 (Commission Decision, 6 April 1994); *Lagerblom v Sweden* App no 26891/95 (ECtHR, 14 January 2003), para 49; *Murtazaliyeva v Russia* App no 36658/05 (ECtHR, 9 May 2017), para 70.

[4] Joshua Karton, 'Lost in Translation: International Criminal Tribunals and the Legal Implications of Interpreted Testimony' (2008) 41 Vanderbilt Journal of Transnational Law 1, 3.

[5] Stefan Trechsel (with the assistance of Sarah J Summers), Human Rights in Criminal Proceedings (OUP 2005) 328.

language and therefore perform its work. It follows that the right to translation and interpretation is a communicative right. The trial is a procedure based on communication, where the defendant has the right to transmit his message to the court. On the other hand, the court must also be capable of receiving this message in order for the communication to be effective. This communicative aspect of the right impacts on certain facets of its implementation, such as the waiving of the right.

The right to translation and interpretation is therefore a procedural, rather than linguistic, right. This is crucial as the ECHR does not aim at protecting or promoting a language or a language-related identity. The native language and linguistic rights of minorities are not protected by Article 6 paragraph 3(e) of the ECHR.[6] The interpretation facilitates communication only for the purpose of safeguarding the fairness of the proceedings. It is not the mother tongue of the suspect (or accused as the case may be) that is protected *per se*. In cases where a defendant is familiar with the language used in court, an interpreter will not be provided (despite his mother tongue being different) as the communication will be deemed successful.[7] Finally, the interpretation can be provided either in his native language or in any other language he speaks or understands.[8]

III. Conditions for Exercising the Right

The presence of an interpreter depends on the linguistic competence of the suspect or accused. However, the measurement of language competence is a daunting task. Even if it is assumed that language skills can be measured, there are other factors that should also be considered in criminal proceedings. Besides the fact that the suspect or accused may not collaborate, making the review of his language skills impossible, the language used in court is a sub-language with great differences from that which a defendant uses in everyday life.

The ECtHR leaves a wide margin of appreciation to Member States to decide whether or not an interpreter should be appointed. According to the ECtHR, a serious examination of the matter suffices.[9] However, in the important case of *Cuscani v UK*, the Court found that the applicant could not follow the proceedings

[6] *Isop v Austria* App no 808/60 (ECHR, 8 March 1962); *K. v France* (1983) 35 DR 203; *Zana v Turkey* ECHR 1997–VII.

[7] *Lagerblom* (n 3), paras 61,62.

[8] *Iovanovski v Republic of Moldova* App no 8006/08 (ECtHR, 5 January 2016), paras 18, 48.

[9] *S.E.K. v Switzerland* App no 18959/91 (Commission Decision, 12 January 1994); *Santa Cruz Ruiz v UK* App no 26109/95 (Commission Decision, 22 October 1997); *Galliani v Romania* App no 69273/01 (ECHR, 10 June 2008), para 54; *Czukowicz v Poland* App no 15390/15 (ECtHR, 24 January 2017), para 20.

despite having the assistance of his brother.[10] The complexity of the charges against the applicant played an important role in the Court's ruling.[11] In recent cases, the ECtHR has held that Member States must take into consideration the linguistic knowledge of the defendant, the nature of the offence with which he is charged, and any communications addressed to him by the domestic authorities to examine whether they were sufficiently complex to require a detailed knowledge of the language used.[12] Ultimately, the trial judge, being the ultimate guardian of the fairness of the proceedings, must examine the matter with scrupulous care.[13]

IV. The Content of the Right

A. The quality of the interpretation

Given that the rights enshrined in the Convention shall be both practical and effective, the competent authorities are obliged not only to appoint an interpreter, but also to examine the adequacy of the interpretation provided.[14] The interpretation must allow the defendant to be informed of the criminal charges against him and defend himself, primarily by presenting his own version of events.[15]

The question then arises regarding whether the defendant should complain about the quality of the interpretation provided in order for that matter to be examined. Indeed, it seems that the ECtHR links the obligation of examining the adequacy of the interpretation by national courts to their proper information about the deficiencies.[16] However, it must be assumed that the information about the deficiencies of the interpretation can come from anyone who is present, as the accused himself is not capable of recognising the errors of the interpreter.[17] Moreover, the court, as the ultimate guardian of the fairness of the proceedings, should evaluate indications of inadequate interpretation *proprio motu*. For instance, in cases where the interpreter does not keep notes while the accused is talking for

[10] *Cuscani v UK* App no 32771/96 (ECtHR, 24 September 2002).

[11] ibid paras 26, 38.

[12] *Hermi v Italy* ECHR 2006–XII 91, para 71; *Güngör v. Germany* App no 31540/96 (ECtHR, 17 May 2001); *Protopapa v Turkey* App no 16084/90 (ECHR, 24 February 2009), para 81; *Katritsch v France* App no 22575/08 (ECtHR, 4 November 2010), para 43.

[13] *Cuscani* (n 10), para 39.

[14] *Kamasinski v Austria* App no 9783/82 (ECHR, 19 December 1989); *Diallo v Czech Republic* App no 20493/07 (ECHR, 28 November 2011), para 23.

[15] *Protopapa* (n 12), para 80; *Katritsch* (n 12), para 42.

[16] *Protopapa* (n 12), para 83; *Katritsch* (n 12) para 42; Stephanos Stavros, The Guarantees for Accused Persons under Article 6 of the European Convention on Human Rights: an analysis of the application of the Convention and a comparison with other instruments (Martinus Nijhoff 1993) 257.

[17] *Priplata v Romania* App no 42941/05 (ECtHR, 13 May 2014), para 93.

a long period of time or transfers a very long answer in a brief, summarised form, the court should be alarmed.

B. The right to an interpeter free of charge

Article 6 paragraph 3(e) of the ECHR clearly states that the accused must be provided with free interpretation. The question that arises is why the accused should be excluded from the interpretation costs, given that in criminal proceedings he is regularly making decisions that have financial repercussions. The answer is simple and demonstrates the specific nature of this right. If the accused, for fear of the financial impact, does not ask for or denies the services of an interpreter, the whole procedure is compromised, as the accused would not be able to understand the procedure nor express his own version of the facts. Thus, while trials may continue without the presence of a defence lawyer, they cannot do so without an interpreter, due to the impact of such an absence on the communication process. Finally, the right to translation and interpretation puts on equal footing an accused who is not conversant with the language of the court and an accused who does speak and understand that language.[18] If there was not such a provision for interpretation free of charge, the accused who is a non-native speaker would be at a disadvantage.

The ECtHR is strict with the implementation of the aforementioned provision. The accused shall not bear the costs in case of either conviction or acquittal.[19] The only instance in which the accused can be charged is in case of abuse of the right. If, for example, he requests the presence of an interpreter unnecessarily or he does not show up to the subsequent trial, he may bear the costs.[20]

C. Translation of documents

The right to translation of documents is not expressly laid down in the Convention. Article 6 paragraph 3(e) of the ECHR speaks of an interpreter and not a translator. However, in accordance with well-established case law, it is clear that the provision covers the translation of documents as well. The accused is entitled to a translation of all those documents or statements in the proceedings necessary to have a fair trial.[21] It would be insufficient to provide the accused

[18] *Luedicke, Belkacem and Koc v Germany* App no 6210/73 (ECHR, 28 November 1978), para 53; *Kamasinski* (n 14), para 75.

[19] *Ozturk v Germany* (1984) Series A no 73, para 57; *Isyar v Bulgaria* App no 391/03 (ECtHR, 20 November 2008), para 45; *Hovanesian v Bulgaria* App no 31814/03 (ECHR, 21 December 2010), para 51.

[20] *Fedele v Germany* App no 11311/84 (Commission Decision, 9 December 1987).

[21] *Luedicke, Belkacem and Koc* (n 18), para 48; *Akbingol v Germany* App no 74235/01 (ECtHR, 18 November 2004).

with only an interpreter, as documents play an important part throughout the proceedings, especially among civil law systems.

Evidently, the accused does not have the right to receive a translation of all the documents in the proceedings. Member States must provide translations of only those documents that are necessary for the accused to benefit from a fair trial.[22] Given that the rules regulating documents that must be translated vary among Member States, it is evident that the burden falls mainly on the defence attorney to request the translation of significant documents beyond those translated by the competent authorities. In addition, it is incumbent on the defence to demonstrate how the absence of a translated document had a negative influence on the fairness of the proceedings. To date, the ECtHR has found no breaches on this ground. Allowing Member States to solely provide the accused with an oral explanation of the necessary documents also demonstrates the relative character of the provision.[23]

V. Waiver of the Right

According to the ECtHR case law, rights enshrined in Article 6 of the ECHR can be waived. Such a waiver, however, cannot run counter to any important public interest.[24] The ECtHR has held that, for a waiver to be effective, it must be voluntary and established in an unequivocal manner.[25] It must also be attended by minimum safeguards commensurate to its importance,[26] such as the full knowledge of the waiver's consequences received by either the court or the defence attorney.[27] Finally, a waiver may be either explicit or implicit.[28]

Nevertheless, some remarks need to be made for cases where an accused does not understand the language of the proceedings. In order for a waiver to be effective, it is required that the accused is capable of understanding the charges laid against him, so that he can properly assess their significance and decide to waive his procedural rights.[29] Moreover, the right to translation and interpretation is more than merely a procedural right. The presence of an interpreter is a prerequisite for a successful communication between the accused and the court. This aspect of the

[22] *Lagerblom* (n 3), para 61.

[23] *Kamasinski* (n 14), para 85; *Husain v Italy* ECHR 2005–III 373; *Baka v Romania* App no 30400/02 (ECtHR, 16 July 2009), para 73; *Katritsch* (n 12), para 41; *Diallo* (n 14), para 23.

[24] Athanasia Dionysopoulou, The Defendant's Right to Examine the Witnesses against him: Article 6 par.3d ECHR (Nomiki Bibliothiki 2017) 112.

[25] *Colozza v Italy* App no 9024/80 (ECHR, 12 February 1985), para 28.

[26] *Poitrimol v France* App no 14032/88 (ECHR, 23 November 1993), para 31.

[27] *Hermi* (n 12), para 79.

[28] *Protopapa* (n 12), para 82.

[29] *Baytar v Turkey* App no 45440/04 (ECHR, 14 October 2014), paras 53–54; *Saman v Turkey* App no 35292/05 (ECHR, 5 July 2011), paras 32, 35.

right affects the possibility of a waiver. Even in case of a waiver, the court can *ex officio* appoint an interpreter to safeguard the fairness of the proceedings.

Furthermore, is must be pointed out that, although most of the rights enshrined in Article 6 of the ECHR are considered defence rights and the accused person and his attorney are being treated as a unity, the right to translation and interpretation cannot be replaced by the presence of a lawyer. The accused has a personal right to be assisted by an interpreter. Therefore, the presence of an attorney alone cannot be regarded as a waiver.

VI. EU Legislation and the ECHR

The regulation of procedural rights by EU law has been strongly debated. The main argument brought forward in this respect is the existence of the ECHR and of the ECtHR that processes and strengthens those rights. Indeed, all Council of Europe Member States are parties to the Convention and new members' compliance to the latter is strictly monitored. A closer examination of the Court's functioning and case law reveals why the protection provided is insufficient.

Firstly, according to Article 35 of the ECHR, there is an obligation to exhaust domestic remedies before lodging an application to the ECtHR. Thus, the protection of the Convention is granted only to those who have the financial means to appear before several domestic courts. Moreover, according to Article 35 of the ECHR, the Court may only deal with an application within a period of six months from the date on which the final decision was taken. This six-month rule diminishes the protection of victims. The potential applicant must, within this temporal limit, search for the appropriate legal representation and consider whether to lodge an application and on which grounds. As a result of the aforementioned restrictions, many violations remain unpunished because of the victim's lack of financial resources.

Furthermore, the inefficiencies of the Strasbourg Court have been revealed since early 2000. The increase in the number of Member States has resulted in an increase in its caseload.[30] It is also important to point out that the judgments are often ineffectively implemented; the Court rarely demands a change in law.[31] The very existence of repetitive cases demonstrates the Court's inability to change the legislative provisions of Member States through its case law. Although the Council, through the Committee of Ministers, monitors the execution of judgments, Member States are primarily responsible for the harmonisation of national legislation with ECtHR case law, having in practice a wide margin of discretion. Unresolved structural problems related to caseload and the lack of a common

[30] Caroline Morgan, 'Where are we now with EU procedural rights?' (2012) EHRLR 427, 430.

[31] See, for example, *Wolfmeyer v Austria* App no 5263/03 (ECtHR, 26 May 2005), para 32.

understanding among Member States on appropriate implementation measures demonstrate the inability of the Court to guarantee rights that are practical and effective.[32]

The European legislative initiative aims to ensure full implementation and respect of Convention standards. Enhancing procedural rights and guaranteeing their consistent application will also ease the burden on the ECtHR. Procedural rights that are important for the defendant will be guaranteed in a uniform manner at national level, leaving other important and innovative issues for the ECtHR.

VII. Directive 2010/64/Eu on the Right to Interpretation and Translation in Criminal Proceedings

The Directive 2010/64/EU[33] (the "Directive") was the first to address issues of procedural criminal justice following the Treaty of Lisbon's entry into force on 1 December 2009. At a general level, it followed the well-established case law of the ECtHR without hesitating to strengthen the protection. In this Part of the article, the aforementioned Directive will be analysed in an effort to highlight its importance and to address its weak points in a mutually beneficial way for both non-native speakers and the courts.

A. The scope of the directive

Article 1 of the Directive refers to the scope of the Directive, regulating both the type of proceedings it applies and its starting point. The Directive applies to criminal proceedings as well as proceedings for the execution of a European Arrest Warrant (EAW). It does not apply to minor cases such as traffic offences. In these cases, if an authority other than a court having jurisdiction in criminal matters has competence for imposing sanctions, it is not required to ensure all the rights provided for in this Directive.[34]

However, the notion of criminal proceedings is not clarified. According to Recital 33 to the Directive's Preamble, the meaning of "criminal proceedings" will be interpreted in light of the case law of the ECtHR, where it is treated as an autonomous concept.[35] As a result, the explicit reference to the EAW execution

[32] *Matthews v UK* ECHR 1999–I 251, para 34; *Čonka v Belgium* ECHR 2002–I 47, para 46.

[33] Directive 2010/64/EU of 20 October 2010 on the right to interpretation and translation in criminal proceedings [2010] OJ L 280/1.

[34] Dir 2010/64/EU, preamble 16.

[35] *Maaouia v France* ECHR 2000–X 273, para 34.

procedure was necessary because extradition procedures are not included in the scope of Article 6 of the ECHR.[36]

The right shall have effect from the moment a person is made aware by the competent authorities that he is suspected or accused of having committed a criminal offence, until the conclusion of the proceedings, namely until *res judicata*. The Directive does not apply after the final determination of guilt. This is also a point where there is compliance with the case law of the ECtHR.[37]

Finally, it must be pointed out that according to Article 1 paragraph 3 of the Directive, in cases of minor offences, where an appeal can be brought before a court having jurisdiction in criminal matters, the Directive shall apply following such an appeal. Whereas this is a deviation from the ECHR, the ECtHR does not differentiate the offences. For determining whether a charge is a "criminal" one, and therefore within the scope of Article 6 of the ECHR, the Court uses the three criteria set out in *Engel*, commonly known as the "Engel criteria".[38] The question naturally arises about whether minor offences that constitute criminal charges according to the ECtHR can be exempted from the scope of the Directive, in view of paragraphs 8 and 32 of the Directive's Preamble.[39]

The scope of the Directive has already been the subject matter of a reference for a preliminary ruling.[40] In *Balogh*, Austria had issued a judgment which, according to the Hungarian law, had to be translated for the purposes of the procedure. According to Hungarian law, in the course of special procedures, the costs of the proceedings are to be paid by the accused if he has been ordered to pay the costs in the main proceedings. The judgment therefore had to be translated not for the protection of the procedural rights of the defendant, but in order for the procedure of recognition of the foreign judgment to be carried out. The question then arose of whether such a procedure can be characterised as a "criminal" one, with the consequence that the defendant shall not bear any costs.

The Court held that the aforementioned procedure could not be considered "criminal" within the meaning of the Directive because the special procedure of recognition of a foreign judgment comes after a final judgment. Therefore, according to Article 1 paragraph 2 of the Directive, it fell outside the scope of the Directive.[41] Moreover, given that Mr. Balogh had received a translated copy of the

[36] *Monedero Angora v Spain E*CHR 2008–IV 429; *Sardinas Albo v Italy* ECHR 2004–I 353.

[37] *Deweer v Belgium* App no 6903/75 (ECHR, 27 February 1980), paras 42, 46; *Eckle v Germany* App no 8130/78 (ECHR, 15 July 1982), para 73.

[38] *Engel and others v the Netherlands* App no 5101/71 (ECHR, 8 June 1976).

[39] See also Debbie Sayers, 'Protecting Fair Trial Rights in Criminal Cases in the European Union: Where does the Roadmap take Us?' (2014) 14 Human Rights Law Review 733, 740.

[40] Case C–25/15 *Proceedings brought by István Balogh* (9 June 2016).

[41] ibid, paras 37–40.

judgment by Austrian authorities, the non-imposition of costs could not be justified in the light of the objectives pursued by the Directive (namely safeguarding the fairness of the proceedings and exercising defence rights).

B. The right to interpretation

According to Article 2 paragraph 1 of the Directive, the services of an interpreter shall be provided to anyone who does not speak or understand the language of the proceedings. This provision applies during police questioning, examination by investigative and judicial authorities, court hearings, and any necessary interim hearings. The interpreter shall be provided without delay, namely within a reasonable period of time.[42]

Moreover, according to Article 2 paragraph 2 of the Directive, the communication between the suspect or accused and his attorney shall be interpreted. Given that until now all the provisions of the Directive grant the same or similar rights with the ECHR, it is evident that this provision was a sticking point. Member States' provisions vary significantly in terms of providing such a right.[43] In some Member States, the communication between the suspect or accused and his attorney is guaranteed in nearly all cases while, in others, restrictions are applied. Cost and time are two key factors that lead Member States to impose restrictions. The defence may sometimes choose to ask for interpretation only to delay the procedure. Furthermore, in some Member States, interpreters are deemed to be in the service of the courts, and therefore should facilitate only the communication between the accused and the court.[44] According to the Directive, the communication between the attorneys and their clients shall be free of charge to safeguard the fairness of the proceedings. The last sentence of Article 2 paragraph 2 of the Directive, however, constitutes an attempt to prevent abuses of the right. The communication shall be interpreted only when it is directly linked to inquiries, hearings, or the lodging of an appeal or other procedural applications. However, when examining the Directive's Preamble, and in particular Paragraphs 19 and 20, it is evident that the same right is there broadly interpreted. An interpreter shall be provided to allow a suspect or accused to explain his version of the events to his attorney, point out any statements with which he disagrees, and make his attorney aware of any facts he wishes to be put forward in his defence. According

[42] Preamble, para 18.

[43] Laurens van Puyenbroeck and Gert Vermeulen, 'Towards Minimum Procedural Guarantees for the Defence in Criminal Proceedings in the EU' (2011) 60 ICLQ 1017, 1034.

[44] Caroline Morgan, 'The New European Directive on the Rights to Interpretation and Translation in Criminal Proceedings' in Sabine Braun and Judith L Taylor (eds), Videoconference and Remote Interpreting in Criminal Proceedings (University of Surrey 2011).

to that interpretation, it is evident that communication between the suspect or accused and his attorney is typically regarded as relevant in the aforementioned circumstances. Most of the time, it will at least include elements falling within the scope of the right, making the presence of an interpreter necessary.

But the question of when the presence of an interpreter is necessary remains. An interpreter is necessary when the suspect or accused is completely unfamiliar with the language used in court. This lack of knowledge of judicial language can be easily traced. In those cases, the appointment of an interpreter also serves the court, because without the presence of a "linguistic intermediary" any contact or communication is impossible. The problem arises when the suspect or accused speaks, or at least has a limited understanding of, the language. How can his language competence be measured to decide whether or not he is capable of following the proceedings and having a fair trial?

First, a basic knowledge of the language used in court is insufficient. The fact that the suspect or accused can communicate in everyday life does not guarantee that he can understand the questions posed to him in their totality or even the charges themselves, considering they are formulated in indecipherable legal language. The language of the court constitutes a sub-language according to sociolinguistics and is vastly different from the one used in other phrasal contexts. The vocabulary is entirely distinct, and even grammar and syntax differ as it is a formal context. Even native speakers have difficulties understanding it. The non-native speaker must therefore be competent enough to understand the jargon used. The nature of the charges plays an important role in this assessment; for example, in a case of theft or defamation, a good knowledge of the language would be sufficient whereas, in cases of complex economic crimes, only a thorough knowledge of the language is acceptable. However, the suspect or accused will often deny to cooperate or deliberately provide incorrect answers. In such cases, the court will have to examine witnesses and documents to evaluate the language proficiency of the suspect or accused. However, this is a time-consuming and expensive procedure prone to failure. In acknowledgement of the inherent difficulties of this procedure, the ECtHR seems to be satisfied with a serious examination of the issue.[45]

The ECtHR takes into account objective criteria, such as the time a person spent in a country, any contact with the educational system, type of work, and the years of residence in a country.[46] While these constitute valuable indicators, they likely will not lead to certainty. The knowledge of a foreign language may require the acquisition of the mechanisms that give productivity and perception, but it does not guarantee an accurate use of the language. It must be also taken into consideration

[45] *Trechsel* (n 5) 334.

[46] *Hermi* (n 12), para 90; *Marzohl v Switzerland* App no 24895/06 (ECHR, 6 March 2012).

that any language difficulties are multiplied in stressful environments such as police departments or courtrooms.[47] Given that the communication between the suspect or accused and the court is also a key factor in guaranteeing the fairness of the proceedings, interpretation shall always be on the suspect's or accused's requests and there are objective indicators that this demand may be justified. The fact that the interpretation is not required to be provided in the native language of the suspect or accused also supports this view. Therefore, if (a) the suspect or accused requests interpretation; (b) there are elements to suggest that his demand is justified; and (c) the interpretation is provided in a language that does not considerably delay the proceedings, an interpreter must be appointed. This solution has the added advantage of resolving the issues related to appointing an interpreter at an early stage in the process. Given that, according to the Directive, an interpreter must be provided during police questioning, who is indeed responsible for deciding at this stage that interpretation is not required and on what grounds? This question cannot wait until the whole issue is examined by the court. If a court assesses for the first time during trial that interpretation is necessary, everything that occurred in pre-trial would be discredited as it can be presumed that the accused was unable to understand not only the procedure, but also the charges against him. Finally, if we also consider that, according to the Directive, communication technology such as videoconferencing, telephone, or the Internet can be used, objections regarding the costs and the delay of the proceedings disappear.

C. The right to translation

The translation of documents in criminal proceedings was also a field where the practice of Member States was disjointed. In some Member States it was not a statutory right whereas in others, provisions varied greatly on what had to be translated. In the Directive, the right to translation is an autonomous right despite not being expressly laid down in the ECHR.

According to Article 3 of the Directive 2010/64/EU, Member States shall ensure that suspects or accused are provided, within a reasonable period of time, with a written translation of all documents essential to enable them to exercise their right to a defence and to safeguard the fairness of the proceedings. Paragraph 2 clarifies the notion of "essential documents" to include at least any decision depriving a person of his liberty, any charge or indictment, and any judgment. During the transposition of the Directive in each Member State, the content of those notions should be specified in accordance with national law. For example,

[47] Ikuko Nakane, 'Problems in Communicating the Suspect's Rights in Interpreted Police Interviews' (2007) 28 Applied Linguistics 87.

it shall be clarified if the notion of a decision depriving a person of his liberty includes only the pre-trial detention or also covers non-custodial alternatives.

The Directive follows the proposal for a Framework Decision put forward by the Commission[48] regarding the explicit reference to the essential documents. There is, however, an important distinction: documentary material is not included in the essential documents that must be translated. Unlike the Commission's proposal, the Directive states that essential documents do not include relevant documentary material such as key witness statements. The Directive resulted from the proposal of thirteen Member States, which in some points offered a narrower protection in comparison to the one drafted by the Commission.[49] Member States opposed the inclusion of documentary material in essential documents that shall be translated, expressing concerns regarding the financial implications of such a provision.[50] Finally, a minimalist version of the provision was adopted, with the documents proposed by the Commission ultimately not being included in the Directive. The position of the United Kingdom was adopted, which emphasised the limited role of documents in common law system.[51] The fact that, in common law systems, there may not even be a "judgment" to be translated possibly explains the use of the word "any" ("*tout*") rather than the definite article "the".[52]

However, access to key documents is important for the suspect or accused to be able to effectively prepare his defence. It is on the basis of the documents constituting the case file that the suspect or accused in civil law systems will call witnesses and gather evidence. Moreover, the knowledge of essential documents can be regarded as serving the sufficient knowledge of the case against him. Therefore, it is difficult to imagine how the rights of the defence can be exercised effectively without the knowledge of important documentary evidence.

Article 3 paragraph 3 of the Directive partially addressed the aforementioned issue. Suspected or accused persons or their legal counsel may submit a reasonable

[48] Commission, 'Proposal for a Council Framework Decision on the right to interpretation and to translation in criminal proceedings' COM (2009) 338 final.

[49] Initiatives of the Member States, 'Initiative of the Kingdom of Belgium, the Federal Republic of Germany, the Republic of Estonia, the Kingdom of Spain, the French Republic, the Italian Republic, the Grand-Duchy of Luxembourg, the Republic of Hungary, the Republic of Austria, the Portuguese Republic, Romania, the Republic of Finland and the Kingdom of Sweden with a view to the adoption of a Directive of the European Parliament and of the Council on the rights to interpretation and to translation in criminal proceedings' (Preparatory acts) 2010/C 69/01.

[50] Steven Cras and Luca de Matteis, 'The Directive on the Right to Interpretation and Translation in Criminal Proceedings' (2010) 4 Eucrim 153, 159.

[51] European Scrutiny Committee, Interpretation and Translation Rights in Criminal Proceedings (HC 2009–10, 5–vii) para 7.16.

[52] James Brannan, 'L'article 3 de la directive 2010/64/UE: la traduction écrite en matière pénale devient un droit à part entire' (2016) 3 Études de linguistique appliquée 281.

request to competent authorities requesting the translation of documents that are essential. Nevertheless, determining whether a document is essential to a specific case is left to the competent authorities, which are not provided with specific guidelines. Such an open provision, leaving a wide margin of discretion for Member States, is hardly aligned with the aim of harmonising the administration of justice between Member States.

The absence of a list of potential essential documents[53] is addressed by the provision granting the right to the suspect or accused to challenge a finding that there is no need for the translation of documents (Article 3 paragraph 5 of the Directive). It must be noted, however, that the suspect or accused must challenge a decision denying the translation of a document to which he does not have access, so as to evaluate its significance. The protection provided relies then on the presence of an attorney who will be capable of evaluating the importance of a document and provide arguments on behalf of the suspect or accused.[54] Moreover, Article 3 paragraph 5 of the Directive provides for a rather vague right to challenge the decision denying the translation of a document. In the Commission's proposal, Member States had to provide a right of appeal against a decision refusing the translation of the documents explicitly mentioned.[55] The Directive, however, includes a right to challenge ("*droit de contester*") in line with the desire of Member States, and in particular of the United Kingdom, not to establish a separate mechanism or complaint procedure for challenging such decisions.[56]

Furthermore, according to Article 3 paragraph 4 of the Directive, there is no requirement to translate passages of essential documents that are not relevant for the purposes of enabling suspects or accused to have knowledge of the case against them. In this way, even the right to translation of the essential documents explicitly mentioned in the Directive is restricted. There is also the possibility to exceptionally replace the written translation of essential documents with an oral translation or oral summary, on the condition that the fairness of the proceeding will not be prejudiced. The oral translation of the essential documents is undoubtedly cost and time-effective. This provision is also in line with the ECtHR's jurisprudence.[57] Nevertheless, it can lead to abusive behavior on the part of Member States. It also deprives the suspect or accused of the possibility to review the relevant

[53] See also Richard Parry, 'Language Rights in Criminal Proceedings and BREXIT: What have we got to lose?' (2017) 2 European Human Rights Law Review 155, 159.

[54] Richard Parry, 'The Curse of Babel and the Criminal Process' (2014) 11 Crim LR 802, 804; Parry (n 53) 159.

[55] COM (2009) 338 final (n 48) Article 3 para 4.

[56] European Scrutiny Committee (n 51).

[57] *Hermi* (n 12), para 70.

information.[58] For those reasons, national provisions should explicitly state the circumstances where an oral translation is exceptionally allowed. The existence of exceptional circumstances and the need to avoid affecting the fairness of the proceedings will undeniably present a challenge for the CJEU.

Covaci was the first time a case was brought before the CJEU in relation to the Directive, after a reference for a preliminary ruling by a German Court.[59] The Amtsgericht Laufen (the local court) had imposed a fine on the accused via a simplified procedure that did not require a hearing or a trial *inter partes*. In this procedure, the decision becomes definite two weeks after its service. The sentenced person is able to lodge an objection securing a trial *inter partes* within the aforementioned time frame. The objection may be lodged either in writing or by a statement recorded by the registry. Mr Covaci lodged an objection in writing using his mother tongue and not German, as he should have done under German law. The question that arose was whether the competent authorities had an obligation to translate the application of Mr Covaci to be admissible and lead to a hearing.

The objection of Mr Covaci cannot be included in the essential documents that should always be translated according to Article 3 paragraph 2 of the Directive. It is not a judgment, a decision depriving a person of his liberty, a charge, or even a document drawn up by judicial authorities. Thus, it remains to be seen if it constitutes an essential document that guarantees the exercise of defence rights and the fairness of the proceedings. National competent authorities are free to evaluate if a document can be characterised as essential, balancing the rights of the suspect or accused with the need to save time and reduce costs. In this case, the CJEU acknowledged that—because the Directive only guarantees minimum rights—Member States are free to expand the list of documents that are considered essential. Therefore, the Court considered the German provision to be compatible with the Directive.

There are, however, a few observations to be made. In this case, there is a distinction drawn between the right to interpretation and the right to translation. According to the CJEU, if Mr Covaci chose to orally lodge a complaint at the registry of the competent court, he would have been provided with an interpreter.[60] Nevertheless, if he had chosen to exercise the same right in writing, a translation would not be provided. There is an inconsistency within the system of the Directive caused by the arbitrary distinction between translation and interpretation. It is an inequality that does not appear to adhere to the spirit and purpose of the

[58] *Brannan* (n 52).
[59] Case C–216/14 *Criminal proceedings against Gavril Covaci* (15 October 2015).
[60] ibid, para 42.

Directive.[61] This inconsistency was also pointed out by Advocate General Bot, who rightly emphasised that, so far as the assistance of an interpreter is guaranteed when an appeal is lodged orally, such assistance must equally be guaranteed where the latter is lodged in writing.[62]

The restriction of essential documents to only three categories of documents drawn up by judicial authorities, and the existence of a wide margin of appreciation in evaluating if a document can be characterised as essential, will certainly raise further issues. A bold jurisprudence by the CJEU, which has to define and elaborate the notion of essential documents, is necessary and will serve as a guideline for Member States. The adoption of a more proactive stance on behalf of the Member States shall also help to ensure that the minimum protection offered by the Directive will trigger a comprehensive protection of the rights of defendants who speak a different language within national provisions.

D. The quality of translation and interpretation provided

According to Article 2 paragraph 8 and Article 3 paragraph 9 of the Directive, the translation and interpretation provided must be of a quality sufficient to safeguard the fairness of the proceedings. The suspect or accused should be capable of exercising his right to a defence and have knowledge of the case against him. Whether the quality of the translation and interpretation provided is sufficient will be judged based on a rather vague concept, such as the suspect's or accused's knowledge (most likely in broad terms) of the case against him and his ability to exercise his right to a defence. Thus, this is a field where subjective judgments prevail. It is also for Member States to decide on the appropriate measures guaranteeing the quality of translation and interpretation. The only condition is that they have to be concrete.[63]

An effort towards the establishment of minimum quality standards for translation and interpretation is made through provisions that demand for the establishment of a register of independent translators and interpreters. Those professionals must be appropriately qualified and must respect the principle of confidentiality. However, the use of the term "independent" is intriguing. In connection with the call to respect the principle of confidentiality, it can be argued that the Directive does not leave room for attorneys or police officers to serve as interpreters/translators. Moreover, the call for respect of the confidentiality principle is the forerunner for a Code of Conduct that will delineate the tasks and establish sanctions in case of malpractice. The respect for the principle of confidentiality would be further reinforced through the creation of a professional

[61] Parry (n 53) 161.
[62] Case C–216/14 *Covaci* (15 October 2015), Opinion of AG Bot, para 80.
[63] Dir 2010/64/EU, Article 5 para 1.

secrecy. Although client-attorney interpretation could be covered by legal privilege,[64] an explicit provision would resolve the issue that goes beyond the client-attorney communication uniformly.[65]

Finally, the Directive introduces an obligation for Member States to ensure that the suspect or accused can complain about the quality of the services provided. In particular, this obligation derives from Article 2 paragraph 5 and Article 3 paragraph 5 of the Directive. But whilst the quality control procedure appears relatively simple in cases of translation (with the appointment, for example, of a second translator who will check the first), how is the quality of interpretation controlled, especially in view of the fact that it is not videotaped or audio-recorded? It could be argued that the suspect or accused has to complain about the quality of the interpretation at the moment it is provided, so as to address this issue quickly and effectively. However, in some cases, the poor quality of the interpretation provided can only be determined *ex post*. In this case the suspect or accused will note obvious divergences between the translated judgment received and the answers that he has given as a result of the interpreter's fault. It is only the obligation to record the interpretation, by audio or video means, that can solve the aforementioned issue and guarantee the fairness of the proceedings.

E. Waiver of the Right

The Directive explicitly states that the suspect or accused may waive the right to translation at any time. According to Article 3 paragraph 8 of the Directive, a waiver of the right to written translation of documents is possible if the person concerned has received prior legal advice or has otherwise obtained full knowledge of the consequences of such a waiver. In line with the case law of the ECtHR, it must also be given voluntarily and in an unequivocal manner.

Such a waiver, however, is not provided for in relation to the right of interpretation. The Directive does not leave any room for waiving this right. This is completely at odds with the jurisprudence of the ECtHR that accepts a waiver in both cases. This choice demonstrates that interpretation is not treated as another defence right, but as a prerequisite for effective communication between the suspect or accused and the court. It is even more important than the presence of an attorney, because the defendant who speaks a different language is in a vulnerable position. The isolation experienced by defendants who cannot understand the language of the proceedings is so severe that it has been compared with patients

[64] *R (Bozkurt) v South Thames Magistrates Court* [2001] EWHC Admin 400.
[65] Parry (n 53) 158.

who have suffered a stroke leading to aphasia; they see that a conversation is taking place, but they cannot participate in any way.[66]

F. Costs

According to Article 4 of the Directive, Member States shall meet all costs of interpretation and translation irrespective of the outcome of the proceedings. This provision is of paramount importance, ensuring that a suspect or accused will not hesitate to ask for an interpreter to fully understand the charges against him and to defend himself.

G. Concluding remarks

The above analysis has demonstrated that there is an extensive use of abstract terms. Key concepts such as "criminal procedure", "essential documents" and "without delay" are not defined in the text of the Directive. Those terms cannot even be interpreted according to national procedural rules, given that there is no express reference to the law of Member States. Those terms could be treated as autonomous concepts.[67] The CJEU has consistently held that a term of a Community law provision which makes no express reference to national provisions for the purpose of determining its meaning and scope must be given an autonomous and uniform interpretation throughout the Community.[68] In the area of procedural rights, the use of autonomous concepts unifies the different national systems, because in every Member State those terms will not be interpreted according to their national arrangements, but according the findings of the CJEU. It also increases the protection for the individual. As Advocate General Bot pointed out, rules adopted on the basis of Article 82(2) of the Treaty on the Functioning of the European Union (TFEU) must be interpreted broadly. Such an interpretation will strengthen the protection of rights as well as mutual trust.[69] In the light of the considerations set out above, the use of general and broad terminology can enhance protections in the area of procedural rights.

[66] Peter Jan Honigsberg, 'Linguistic Isolation: A New Human Rights Violation Constituting Torture, and Cruel, Inhuman and Degrading Treatment' (2014) 12 Northwestern University Journal of International Human Rights 1.

[67] Valsamis Mitsilegas, EU Criminal Law after Lisbon: Rights, Trust and the Transformation of Justice in Europe' (Hart Publishing 2016) 178.

[68] Case C–327/82 *Ekro BV Vee-en Vleeshandel v Produktschap voor Vee en Vlees* [1984] ECR I–107, para 11; Case C–195/06 *Kommunikationsbehörde Austria (KommAustria) v Österreichischer Rundfunk* (ORF) [2007] ECR I–08817, para 24; Case C–316/05 *Nokia Corp. v Joacim Wärdell* [2006] ECR I–12083, para 21.

[69] *Covaci* (n 59) Opinion of AG Bot, para 33.

VIII. Concluding Remarks on the Case Law of the ECtHR and the Directive's Provisions

The Directive on the right to translation and interpretation undoubtedly followed the case law development of the right by the ECtHR. Article 32 of the Directive's Preamble explicitly states that the level of protection should never fall below the standards provided by the ECHR as interpreted in the case law of the ECtHR. Accordingly, Article 8 of the Directive states that nothing in the Directive shall be construed as limiting or derogating from any of the rights and procedural safeguards ensured under the ECHR. Thus, the Directive at least confers the same rights as the ECHR, with the important difference that it imposes them directly on Member States. The Directive does not only impose a mere *ex post* control, but Member States are invited to amend their national provisions. Outlined below are the provisions where the Directive followed the case law of the ECtHR and the ones that provide greater protection.

The Directive gave regulatory autonomy to the right of translation. Although the enshrining of the right, with some restrictions, reflects the necessary adjustments made at negotiations, the fact that the right is explicitly guaranteed constitutes a great difference. Moreover, in line with the case law of the ECtHR, the Directive states that all costs should not be borne by the defendants. The protection of this financial aspect of the right is absolute both in the Directive and the ECHR, and is a feature of the right to translation and interpretation on which no restrictions can be imposed. In all other aspects of the right, restrictions can be imposed. In the ECtHR's case law, such restrictions are examined *ad hoc* as part of an overall approach. The Directive also leaves a margin of discretion to Member States in establishing the mechanism that will define who requires the services of an interpreter. Furthermore, although it is stated that the interpretation provided must be of sufficient quality to safeguard the fairness of the proceedings, the decision is left to national courts (considering whether the suspect or accused had knowledge of the charges against him and was capable of defending himself). In relation to the right of translation, Member States shall provide a translation of all essential documents. However, there is a great difference with documents explicitly referred to in Article 3 paragraph 2 of the Directive (*i.e.* judgments), whose protection is absolute. For example, in case of the translation of a judgment, there is no room for weighing in order for the competent authorities to conclude if there was a breach of defence rights. In the case of any other document that may be considered "essential", but it is not explicitly mentioned in Article 3 paragraph

2 of the Directive, the competent authorities have the opportunity to weigh the costs and the delay involved with the fairness of the proceedings.

Guaranteeing the right of interpretation in relation to attorney-client communication is also an innovation introduced by the Directive. An effort to establish a minimum quality of interpretation services has also been made through administrative measures (Articles 5 and 6 of the Directive).

Nevertheless, the vague terminology used and the wide margin of discretion left to Member States is a daunting issue. Therefore, the transposition of the Directive and the approach of the CJEU will be of paramount importance.

Data Catalysis, Informational Violence, and the Denaturalisation of the Natural Person

Asad Rizvi*

I. The New Age of Discovery

Beneath the masts of the new conquistadors stretch the uncharted high seas of the digital where no state is able to claim jurisdiction. Here, hegemonic merchants circumnavigate between the Old World and the New to trade in bullion of unprecedented properties: data. Across this wild expanse, private actors have speared their flags of domination, governed only by commodious mercantile treaties. Laws that purport to protect local populations jostle powerlessly against an amassing tide of state-sanctioned yet private monolithic power, steered hypnotically by standard-bearers of a civilising faith to which dutiful observation is widely deemed a noble calling for humankind.[1]

Today, the dominant 'civilising missions' involve no priests. Cyberspace has appeared as the Earth's freshest *terra nullius* with a population undergoing an edification of its own kind. With the 'Death of God'[2] and the subsidence of religions as universal guiding truths during the latter half of the last century, market

* LL.B. Law (First Class Hons) and current LL.M. Human Rights Student at Birkbeck College, University of London. The author would like to express gratitude to Dr Oscar Guardiola-Rivera and Professor Bill Bowring, to whom this paper was originally submitted, and whose guidance was inspirational. The author is also deeply indebted to the *Cambridge Law Review* board for the opportunity, and to family members for patience and their proofreading of drafts.

1 Claims to new territorial sovereignty over lands deemed ungoverned by Europeans, succeeded only through the 'civilising missions' of the Churches, justified on grounds of natural reason. See John Witte and Richard C Martin (eds), *Sharing the Book: Religious Perspectives on the Rights and Wrongs of Proselytism* (Wipf and Stock 2008) 163.

2 Friedrich Wilhelm Nietzsche and Walter Arnold Kaufmann, *The Gay Science: With a Prelude in Rhymes and an Appendix of Songs* (1st edn, Vintage Books 1974) 167, 181.

efficiency has emerged as the primary metanarrative of power legitimation.[3] Just as some indigenous populations first received Christianity as a complement rather than a threat to their own belief systems,[4] online sociality has emerged as a virtual but independent prosthetic to the 'real',[5] whilst concealing mercantile functionalities under veneers of leisure.[6] Across the digital network, mechanisms of economic betterment prevail to steer happiness to such extent that humankind is in the process of handing over the helms of reasoned judgment to automated apparatuses of efficiency.[7] Human experience hence is distilled into quantified algorithmic input to improve shared understanding and to maximise revenue.[8]

[3] In 1984, Jean-François Lyotard presupposed the data revolution when he wrote that, through science, the contemporary age could be defined by "the incredulity toward metanarratives". "The decision makers", he wrote,

> attempt to manage these clouds of sociality according to input/output matrices, following a logic which implies that their elements are commensurable and that the whole is determinable. They allocate our lives for the growth of power. In matters of social justice and scientific truth alike, the legitimation of that power is based on its optimizing the system's performance—efficiency.

See Jean-François Lyotard, *The Postmodern Condition: A Report on Knowledge* (Geoff Bennington and Brian Massumi trs, University of Minnesota Press Minneapolis 1984) xxiv.

[4] Avelar relates that, in the 16th and 17th centuries, the Tupinambá people of present-day Brazil

> appeared malleable, accepting, and mimetic of the Portuguese values only, in a second moment, to look like they had forgotten everything and moved on to something else. In other words, what stunned the Portuguese was not the fact that there was a completely different set of beliefs in play. It was not the presence of a cosmogony contradictory with the Christian one. It was, rather, that the Tupinambá seemed to operate outside the Aristotelian logic of identity and non-contradiction altogether.

See Idelber Avelar, 'Amerindian Perspectivism and Non-Human Rights' (2013) 1 Alter/Nativas, 11–12.

[5] This can be epitomised by John Perry's *Declaration of the Independence of Cyberspace*, of 8 February 1996: "Governments of the Industrial World, you weary giants of flesh and steel, I come from Cyberspace, the new home of Mind. On behalf of the future, I ask you of the past to leave us alone. You are not welcome among us. You have no sovereignty where we gather." See Aron Mefford, 'Lex Informatica: Foundations of Law on the Internet' [1997] Indiana Journal of Global Legal Studies 211, 218.

[6] Adorno and Horkheimer's 1944 assertion that "[e]ntertainment is a prolongation of work under late capitalism" resonates ever more prevalently in the context of social media." See Theodor W Adorno and Max Horkheimer, 'Culture Industry: Enlightenment as Mass Deception' in *Dialectic of Enlightenment* (Blackwell Verso 1997) 109.

[7] For example, in 2017, Mark Zuckerberg laid out a manifesto for Facebook where AI fights for a "common understanding", identifies "risks" and decides which facts can be deemed credible or not. He talks of making leaps "from tribes to cities to nations" and reaching the next level of social infrastructure primarily through automation. See Mark Zuckerberg, 'Building Global Community' (6 February 2017) <https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10103508221158471/?pnref=story> accessed 21 May 2018.

[8] Rieder notes, "[s]oftware, again, is used to formalize and disambiguate notions of value and the resulting value signals are both directed to market participants and ranking algorithms. See Bernhard Rieder, 'Beyond Surveillance: How Do Markets and Algorithms "Think"?' (2017) 3 Le Foucaldien, 7.

As network and processing speeds increase with limitless storage possibilities, data technologies now are able to process previously inconceivable quantities of information to advance human knowledge and experience.

A paradox of the post-war era is that the globalisation of data technology has exploded in tandem with the most active period of human rights development. For Moyn, the term 'human rights' evokes romantic ideals of global utopia constructed upon individual dignity.[9] Yet he observes that the slipstream of market determinism has endorsed the simultaneous decline of social and economic rights, driving the betterment of individual wealth before that of individual well-being.[10]

The application of human rights has diversified since the Universal Declaration of Human Rights 1948 (UDHR), and so too have modes of power which have learned to circumvent its core principles. With new technologies proliferating modes of state oppression, many more subtle forms of violence remain unrecognisable within the original human rights framework, with latent potentialities for unprecedented consequences. Although 'traditional' weaponry remains in very much in use by states and their forces, more subtle disciplinary apparatuses—or *dispositifs*, as Foucault would call them[11]— are emerging through the processing of personal and bulk data in the private and public sphere. International legislators are not blind to these emerging forms of violence, but their ethereal properties make them uncontainable within traditional legal boundaries. This article seeks to address these leakages with a view to proposing a more comprehensive framework than is currently available. Must the jurisprudence of human and fundamental rights continue solely to restrain itself to the refrain of 'never again', or can it take another step forward with a more preventative objective of 'never shall it be'?

II. A Macro-Evolution of the Micro

A. Data catalysis

Numerous modes of data transformation exist as a result of the application of algorithmic processes upon user information, but since there is no common terminology for these disparate processes, this paper will refer to the collective paradigm as 'data catalysis'. Much like chemical catalysis,[12] algorithms increase the rate of forward and reverse transformations, and the energy threshold required is

[9] Samuel Moyn, *The Last Utopia: Human Rights In History* (Belknap Press of Harvard University Press 2012)

[10] ibid 35–36.

[11] Michel Foucault, *The History of Sexuality* (Vintage Books 1990) 96, 140.

[12] Antony Spiers and Derek Stebbens, *Chemistry by Concept* (Heinemann Educational Books 1973) 142; Donald A McQuarrie *et al, General Chemistry* (4th edn, Univ Science Books 2011) 663.

lowered, enabling large quantities to be processed at speed. In data catalysis, the 'catalyst' is algorithmic AI, and the 'substrate' can be thought of as the user's data. Automation accelerates processes of data transformation that otherwise would have to be untaken manually.

Within this umbrella term are contained numerous linked but distinct phenomena, including data mining,[13] personal data,[14] metadata,[15] big data,[16] artificial intelligence (AI),[17] behavioural targeting,[18] and algorithmic decision-making.[19] The processes commence with material acquisition and complete with transformation. Although there is a broad range of applications for data catalysis across multiple disciplines, it is the efficacy sought over our own species that simultaneously has rendered terms like 'metadata' and 'big data' to become both buzzwords and profanities in the media of recent years.

Clearly, the issue of most conspicuous controversy for data catalysis is that of privacy. This article would do little to add to the already voluminous body of

[13] The Oxford Dictionary defines 'data mining' as: "the practice of examining large pre-existing databases in order to generate new information." See Angus Stevenson (ed), *Oxford Dictionary of English* (3rd edn, OUP 2010).

[14] Article 4(1) of GDPR defines 'personal data' as:

> any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

See General Data Protection Regulation, 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46' (2016) 59 Official Journal of the European Union (OJ) 294.

[15] The Oxford Dictionary defines 'metadata' as: "[a] set of data that describes and gives information about other data". See: Stevenson (n 13).

[16] The Oxford Dictionary defines 'big data' as: "[e]xtremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions". See ibid. A full definition is given in the next section.

[17] The Oxford Dictionary defines 'artificial intelligence' as: "[t]he theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages". See ibid.

[18] Blue Fountain Media define 'behavioural targeting' as:

> a technique used by online publishers and advertisers to increase the effectiveness of their campaigns through information collected on an individual's Web-browsing behavior, such as the pages they have visited or the searches they have made, to select which advertisements to display to that individual. The technique helps deliver online advertisements to the users who will be the most interested in them. Behavioral data can also be combined with other user information such as purchase history to create a more complete user profile.

See Blue Fountain Media, 'Behavioral Targeting – Glossary' (*Blue Fountain Media*, 29 October 2011) <https://www.bluefountainmedia.com/glossary/behavioral-targeting/> accessed 8 January 2018.

[19] The automated decision-making by computer algorithms, linked to AI.

literature on personal data and traditional surveillance. The issues surrounding metadata and big data, however, are not entirely detached, and must be understood together within the context of personal information.

Over the course of the post-war period, the Orwellian narrative has awakened the public to the diversity of personal intrusions that every new form of communication technology carries. Notwithstanding the ever-changing manner in which privacy violations take place, the principles established in Article 12 of the UDHR (which alludes to non-interference with an individual's privacy, family, home or correspondence) and Article 8 of the European Convention of Human Rights (ECHR) (right to respect for private and family life) are inescapable where the content of private information is intercepted, including those of proportionality against competing interests.[20] In Europe, the legal right to privacy has flowed with relatively healthy correlation to developments in information technology.[21] Rooted in Article 8 ECHR privacy rights, and the European Union (EU) Charter's derivative right to data protection,[22] personal data remains the remit of the individual.[23]

Conversely, federal laws in the United States of America (US) do not provide similar guarantees for the individual, favouring the interests of national

[20] Such as the limitations in Article 8(2), or where there is a 'reasonable expectation of privacy' as per *Campbell v Mirror Group Newspapers Ltd* [2004] UKHL 22.

[21] In 1980, the OECD was quick to recognise issues of inter-jurisdictional data flow by setting out its Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, a non-binding guidance. These principles were adopted by the Council of Europe a year later in Convention 108 to set down limitations in how data is handled whilst maintaining a consistent flow of data for the purposes of trade. The 1995 Data Protection Directive (DPD) served to ratify the objectives of Convention 108 at European Community level so as to ensure a harmonised protocol for both automated and non-automated data across both the public and private sectors. Although the Directive set out to protect individuals' Article 8 rights, it made no mention of human rights, instead focusing on the procedural duties of data controllers. In 2000, the Charter of Fundamental Rights of the European Union first established data protection as a right unto its own, thus consolidating the principles set out in the 1995 DPD and ECHR. The Charter did not come into force until 2009. See European Convention for the Protection of Human Rights and Fundamental Freedoms 1950; Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data 108 European Treaty Series (1981); Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data 1995; Charter of Fundamental Rights of the European Union 2000; Sian Rudgard, 'Origins and Historical Context of Data Protection Law' in Eduardo Ustaran and International Association of Privacy Professionals (eds), E*uropean Privacy: Law and Practice for Data Protection Professionals* (International Association of Privacy Professionals 2012) 6–17.

[22] *Charter of Fundamental Rights of the European Union.*

[23] *Copland v The United Kingdom* No. 62617/00 (European Court of Human Rights 4 March 2007).

security and law enforcement.[24] Where data protection exists for US citizens, it usually does not apply to foreign nationals.[25] In 2000, the EU Commission's *Safe Harbour* decision declared the exchange of data between the EU and US to be consistent with the provisions of the EU's 1995 Data Protection Directive, and that US data protection laws in the US sufficed to provide equivalent protection for Europeans.[26] Yet the *Safe Harbour* accord did not preclude government agencies in the US from indiscriminately accessing EU residents' personal data amidst the wave of post-9/11 emergency legislation.[27] These included a 2008 amendment to grant immunity to private firms that offered assistance to intelligence agencies,[28] thereby creating a corporate buffer for state surveillance. The National Security Agency (NSA) itself boasted "direct access" to the servers of nine major consumer companies—including Microsoft, Yahoo, Google, Facebook, and Apple—as part of its PRISM surveillance programme.[29]

Back across the ocean, the British Government Communication Headquarters' (GCHQ) Project Tempora programme placed interceptors on 200 of the underwater cables that came to shore, collecting 21 petrabytes of data per day[30], all of which was shared with the NSA and its 850,000 private contractors.[31] On account of the United Kingdom's (UK) advantageous geographic location between Europe and North America, GCHQ were able, by 2010, to intercept a quarter of the world's internet traffic,[32] making it arguably the most extensive and

[24] Policy Department C: Citizens' Rights and Constitutional Affairs, 'A Comparison between US and EU Data Protection Legislation for Law Enforcement Purposes' (European Parliament (Directorate General For Internal Policies) 2015) 7.

[25] ibid.

[26] *Commission Decision 2000/520/EC 2000* [2000] L 215/7 OJ.

[27] These include: Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act (2001) 107–56 (United States); These include: Homeland Security Act (2002) 107–296 (United States); Detainee Treatment Act (2005) 109–148 (United States); Military Commissions Act (2006) 109–366 (United States); Foreign Intelligence Surveillance Act of 1978 Amendments Act (2008) 110–261 (United States).

[28] *Foreign Intelligence Surveillance Act of 1978 Amendments Act (2008).*

[29] Paul Bernal, 'Data Gathering, Surveillance and Human Rights: Recasting the Debate' (2016) 1 Journal of Cyber Policy 243, 4.

[30] The equivalent of 30.5 million CD-ROMs per day: Tim Fisher, 'Terabytes, Gigabytes, & Petabytes: How Big Are They?' (Lifewire, 20 September 2017) <https://www.lifewire.com/terabytes-gigabytes-amp-petabytes-how-big-are-they-4125169> accessed 12 December 2017.

[31] Kadhim Shubber, 'A Simple Guide to GCHQ's Internet Surveillance Programme Tempora' (WIRED UK, 24 June 2013) <http://www.wired.co.uk/article/gchq-tempora-101> accessed 12 December 2017.

[32] GCHQ, 'Supporting Internet Operations' (GCHQ 2010) 3.

intrusive intelligence agency of the 'Five Eyes' group of nations comprising the US, UK, Canada, New Zealand, and Australia.[33]

B. 'This is just metadata'

The scale of the NSA's and GCHQ's surveillance programmes came to light in 2013 with the Edward Snowden leaks, revealing major flaws as a result of jurisdictional disparities, as well as new and unrecognised forms of surveillance.[34] The reassurance by Dianne Feinstein, chair of the Senate intelligence committee, that "this is just metadata"[35] did little to quell the fears of data experts who already understood its implications.[36] Although not as immediately invasive as phone-tapping, Bernal notes that metadata is more efficient for surveillance than content.[37] Snowden stated that GCHQ collected metadata from "every visible user on the Internet".[38] Despite the arduous and unreliable task of filtering high volumes of data, the complexity of collected metadata could reveal a plethora of personal details and relationships;[39] and as the former head of the CIA, General Michael Hayden, candidly stated, "we kill people based on metadata".[40] Although most citizens will not suffer assassination as a result of e-mail headers, the very existence and public knowledge of modern panoptic technologies, as Richards observes, has

[33] Ewen MacAskill et al, 'Mastering the Internet: How GCHQ Set out to Spy on the World Wide Web', *The Guardian* (21 June 2013) <http://www.theguardian.com/uk/2013/jun/21/gchq-mastering-the-internet> accessed 13 December 2017.

[34] Bernal (n 29) 4–5.

[35] John Naughton, 'NSA Surveillance: Don't Underestimate the Extraordinary Power of Metadata', *The Guardian* (21 June 2013) <http://www.theguardian.com/technology/2013/jun/21/nsa-surveillance-metadata-content-obama> accessed 12 December 2017.

[36] Matt Blaze, 'Phew, NSA Is Just Collecting Metadata. (You Should Still Worry)' (*WIRED*, 19 June 2013) <https://www.wired.com/2013/06/phew-it-was-just-metadata-not-think-again/> accessed 13 December 2017.

[37] Bernal (n 29) 6.

[38] Nigel Morris, 'Edward Snowden: GCHQ Collected Information from Every Visible User on the Internet' (The Independent, 25 September 2015) <http://www.independent.co.uk/news/uk/home-news/edward-snowden-gchq-collected-information-from-every-visible-user-on-the-internet-10517356.html> accessed 12 December 2017.

[39] Shubber (n 31).

[40] Johns Hopkins University, 'The Johns Hopkins Foreign Affairs Symposium Presents: The Price of Privacy: Re-Evaluating the NSA' (7 April 2014) <https://www.youtube.com/watch?v=kV2HD-M86XgI> accessed 12 December 2017.

a substantially disfiguring effect on power dynamics and threatens the effective functioning of democracy.[41]

The Snowden revelations caused an Austrian citizen, Maximillian Schrems, to take a case to European Court of Justice that, in 2014, struck down the entirety of the *Safe Harbour*[42] agreement for its inadequacy in safeguarding EU residents' rights against indiscriminate surveillance by US government agencies.[43] Citing the *Digital Rights Ireland*[44] case, the Court emphasised that the acquisition of information is enough to establish an interference of rights.[45]

In 2016, shortly after the Schrems decision, the EU finalised the General Data Protection Regulation 2016 (GDPR) that came into force in May 2018. The regulation constitutes the most comprehensive piece of data protection legislation seen globally to date, and sets out a number of developing rights including the right to be forgotten, the right to restrict processing, and the right to be informed.[46] Although, as van der Sloot reflects, these rights are constructed as fundamental rights that apply horizontally between private parties,[47] they originate from the right to privacy.[48] All such rights concern the individual, but leave exposed many other issues that emerge across the data catalysis paradigm.

III. Micro-Devolutions of the Macro

Big data's relationship to the individual is less immediate. Whilst there is no agreed definition of 'big data', a common conception is that which comprises the 'four Vs':[49] the collection of large *volumes* of data, from *various* sources, processed

41 Neil M Richards, 'The Dangers of Surveillance' (2013) 126 Harvard Law Review 1934, 1951–1952.

42 *Commission Decision 2000/520/EC* (n 26).

43 Case C–362/14 *Maximilian Schrems v Data Protection Commissioner* [2015] ECLI:EU:C:2015:650.

44 Case C–293/12 *Digital Rights Ireland and Seitlinger and Others* [2012] ECLI:EU:C:2014:238.

45 "To establish the existence of an interference with the fundamental right to respect for private life, it does not matter whether the information in question relating to private life is sensitive or whether the persons concerned have suffered any adverse consequences on account of that interference (judgment in *Digital Rights Ireland* and Others, C–293/12 and C–594/12, EU:C:2014:238, paragraph 33 and the case-law cited)." See *Maximilian Schrems v Data Protection Commissioner* (n 43) para 87.

46 *Regulation* (n 14).

47 Bart van der Sloot, 'Legal Fundamentalism: Is Data Protection Really A Fundamental Right?' in Ronald Leenes et al (eds), *Data Protection and Privacy: (In)visibilities and Infrastructures*, vol 36 (Springer 2017) 6–7, 13.vol 36 (Springer 2017

48 ibid 5–6.

49 Bart van der Sloot et al (eds), *Exploring the Boundaries of Big Data* (Amsterdam University Press 2016) 14; IBM, 'Infographic: The Four V's of Big Data | IBM Big Data & Analytics Hub' (*IBM Big Data and Analytics Hub*, no date) <http://www.ibmbigdatahub.com/infographic/four-vs-big-data> accessed 30 November 2017.

at high *velocity*, and checked for *veracity*.[50] Once bulk data has been acquired and processed it is used to create predictive crowd data, and applied through profiling.[51] Big data, as Oostveen points out, is far from a unitary phenomenon, but a set of processes that invoke separate legal issues at every stage.[52]

Before addressing the widening concerns on subsequent pages, it is important to point out that blanket condemnation of these new technologies might be premature as big data and AI have the capacity to substantially assist human rights. Professors McGregor and Walden have pointed out the value of data analysis in monitoring and discerning patterns in human rights abuses, in establishing the accountability of perpetrators, and even in identifying human bias and preventing discrimination.[53]

Notwithstanding these genuinely positive applications of big data, the threats the new technology can pose to human rights are far more cogent, and require swift and perceptive legal responses to ensure advances do not run amok. Whilst European data protection law is becoming ever more sophisticated in response to technological demands, the 'traditional' privacy framework leaves open ambiguities.

Whereas personal data acquired through surveillance practices would fall under privacy laws, anonymised data circumvents the same laws for lack of identifiability.[54] Where the GDPR considers identifiable personal data to be the property of natural persons, it detaches the user from her data after anonymisation, with the controller retaining rights over the database in intellectual property law.[55] Where private data surveillance seeks to ascertain individuals' details, big data primarily seeks out crowd trends whereupon it constructs predictions.[56] Whereas private data can be located on a single computer system, big data is dispersed across a network or multiple networks, its multi-nodal nature often storing information in numerous jurisdictions, exposing it to unconsented access. Whereas nodes of information can easily be linked together in traditional data sets, big data relationships are more voluminous and are complex to ascertain.[57] Although the

[50] Laux also proposes two more Vs: the legal *validity* of data in hand, and *volatility* of changes in the world that might affect its relevance. See Christian Laux, 'The Legal Aspects of Big Data' [2014] Swiss Analytics Magazine, 15–16.

[51] Sloot *et al* (n 49) 9.

[52] Manon Oostveen, 'Identifiability and the Applicability of Data Protection to Big Data' (2016) 6 International Data Privacy Law 299, 300–302.

[53] Bingham Centre for the Rule of Law, 'Artificial Intelligence, Big Data and the Rule of Law' (Event Report, The Law Society 9 October 2017) 4, 7.

[54] Oostveen (n 52) 306–307.

[55] Richard Kemp et al, 'Legal Rights in Data' (2011) 27 Computer Law & Security Review 139, 2.

[56] Oostveen (n 52) 301–302.

[57] Deepali Aggarwal, 'Difference between Traditional Data and Big Data' (Project Guru, 30 June 2016) <https://www.projectguru.in/publications/difference-traditional-data-big-data/> accessed 30 November 2017.

appropriation of personal data is clearly subject to transparency principles under data protection law, the use of that information once anonymised is less defined.

As the Snowden leaks revealed, not only are states able to evade their data protection duties behind the shield of private companies,[58] but non-state parties have also created a space where private actors wield a similar degree of influence and power over individuals' lives with scant accountability.[59] The Internet is the only mode of human communication that remains unregulated by a binding international treaty[60] in key with the neo-liberal fondness of informal and tractable forms of discretionary law.[61] Zalnieriute observes that, on account of the private stewardship on the net, companies have set out their own thresholds of human rights enforcement in accordance with their commercial goals.[62] The ensuing empirical effect on individuals is that of an unprecedented form of social contract without the need to establish formal state sovereignty. As Mejias analogises, the relationship of users to the digital network is reminiscent of colonialism, whereby colonial power imposed subjecthood without offering citizenship.[63] After the

[58] Another example is of the US telecommunication provider Verizon being ordered to hand over details of all calls to the NSA. See *In re application of the Federal Bureau of Investigation for an order requiring the production of tangible things from Verizon Business Network Services, Inc on behalf of MCI Communication Services, Inc d/b/a Verizon Business Services* No. BR 13-80 (United States Federal Foreign Intelligence Surveillance Court 25-4-13).

[59] To exemplify, Mark Zuckerberg's hearing at the Senate, not taken under oath, was marked by interactions of this sort:

> [SENATOR] FLAKE: …[D]o you believe that Russian and/or Chinese governments have harvested Facebook data and have detailed data sets on Facebook users? Has your forensic analysis shown you who else, other than Cambridge Analytica, downloaded this kind of data?
>
> ZUCKERBERG: Senator, we have kicked-off an investigation of every app that had access to a large amount of people's data before we locked down the platform in 2014. That's underway, I imagine we'll find some things, and we are committed to telling the people who were affected when we do. I don't think, sitting here today, that we have specific knowledge of—of other efforts by—by those nation-states. But, in general, we assume that a number of countries are trying to abuse our systems.

See Bloomberg Government, 'Transcript of Mark Zuckerberg's Senate Hearing', *Washington Post* (10 April 2018) <https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/> accessed 8 June 2018.

[60] Monika Zalnieriute, 'The Anatomy of Neoliberal Internet Governance: Queer Critical Political Economy Perspective' in Dianne Otto (ed), *Queering international law: possibilities, alliances, complicities, risks* (Routledge research in international law, Routledge 2018) 15.

[61] William E Scheuerman, 'Economic Globalization and the Rule of Law' (1999) 6 Constellations 3.

[62] Zalnieriute (n 60) 26.

[63] Ulises Ali Mejias, *Off the Network: Disrupting the Digital World* (Electronic Mediations Volume 41, University of Minnesota Press 2013) 8.

opaque processes of anonymisation, as discussed in the next section, subtle forms of violence begin to occur within the indeterminate zone of the digital.

IV. Into the Big Data Processing Plant

The procedures of big data are diverse and complex. To correctly assess the interplay between human rights law and big data, we must look at its processes rather than effects. Oostveen has identified a consolidated model to simplify its processes into three overarching phases—namely, acquisition, analysis and application[64]—that can assist us in identifying human rights issues at every stage.

A. Mining the Raw Materials

At the first stage is the *acquisition* of big data's raw materials.[65] These can be identifiable and anonymous data, as gathered through data mining, consensual data disclosure (such as through social media data), data sensors (such as global positioning system (GPS)), surveillance, and from the sale of data to third parties.[66] McDermott notes that, irrespective of whether activity is taking place in solitude, between users, or between users and above, surveillance is being conducted continually from the watchtowers of state or private institutions, whether by human or machine interception.[67] Data appropriation at this stage relates to traditional forms of raw and unprocessed monitoring, in multiple formats, of individuals in terms of their *natural* personhood. As such, standard Article 8 privacy rights apply, as does the fundamental right to data protection, and its subsidiary rights. Article 4(1) of the GDPR defines personal data as being "any information relating to an identified or *identifiable natural person*",[68] listing a number of identifiers such as name, online handles, location data, or other personal traits that are now familiar within human rights instruments.

B. Alchemic refinements

After the data is acquired, it must then be analysed.[69] It is at this stage that big data's new paradigms emerge. As big data is concerned with trends, data sets are assimilated and often anonymised.[70] On paper, anonymised data ought to fall

64 Oostveen (n 52) 306.
65 ibid 306–307.
66 ibid 307.
67 Yvonne McDermott, 'Conceptualising the Right to Data Protection in an Era of Big Data' (2017) 4 Big Data & Society, 4.
68 Emphasis added.
69 Oostveen (n 52) 307.
70 ibid 301.

below the threshold for data protection. What remains opaque is the process that follows anonymisation.

In 2006, Netflix announced a public contest to offer a prize for best film recommendation algorithm by releasing a data set containing 500,000 anonymised film recommendations.[71] Researchers Narayanan and Shmatikov revealed the 'leakiness' of this dataset by cross-correlating it against users' publicly available film ratings on the Internet Movie Database (IMDb) and were able to identify individual records and even ascertain political, religious and sexual preferences.[72]

When two or more anonymised datasets are combined, it is therefore possible to merge data to produce 'commingled data',[73] which too can compromise individual privacy.[74] The process could be likened to a notepad of tracing paper in which indistinct facial features are drawn on every page, but when overlaid with other pages, a clear identity can be ascertained. Data sources are growing by the day and include government, commercial, transactional, private, open-source, electoral and lifestyle, among other forms of datasets,[75] many of which are available to all as public or open data.[76] Given the relational nature of big data, information can also be gathered about a person from the data that is mined from others they know.[77] The technical possibility of irreversible anonymisation is widely refuted, opening up the risk of misuse by third parties.[78] Further concerns have been raised about AI systems choosing data sources by themselves, and creating metadata through their own analysis, for example facial recognition software being used to take guesses at a user's sexuality,[79] thereby taking liability for discriminatory analysis away from human actors.

Even where data is legally compliant, analysis can produce discriminatory information, unbeknownst to users. Data protection rights will apply to the new data, but traceability is difficult where companies do not follow principles of

[71] Kate Greene, 'The $1 Million Netflix Challenge' (*MIT Technology Review*, 6 October 2006) <https://www.technologyreview.com/s/406637/the-1-million-netflix-challenge/> accessed 5 September 2018.

[72] Arvind Narayanan and Vitaly Shmatikov, 'Robust De-Anonymization of Large Sparse Datasets (How to Break Anonymity of the Netflix Prize Dataset)' in (2008 IEEE Symposium on Security and Privacy, IEEE May 2008) 11.

[73] Kemp *et al* (n 55) 29–30.

[74] Oostveen (n 53) 307.

[75] Graham Smith, 'How To Build Geodemographics From Big Data' (CACI March 2016) 13.

[76] ibid 17; Great Britain and others, *Open Data White Paper: Unleashing the Potential.* (Stationery Office 2012) 8.

[77] McDermott (n 67) 4.

[78] Oostveen (n 52) 306.

[79] Bingham Centre for the Rule of Law (n 53) 5.

transparency.[80] For all its positive steps forward, the GDPR lays down no binding terms on commingled data, only paying it lip service in Recital 26 of the Preamble.[81] Its force applies to "identifiable natural persons", including pseudonymised data that can be re-identified.[82] It goes on to clarify, however, that data protection does not apply to anonymous information.[83] The legal personhood of the natural person, once de-identified, is dissolved within the digital space. Nature, by this jurisprudence, is contingent on identity. Therefore, when stripped of identity, a person is also stripped of natural personhood.

C. Into the turbines of power

Big data's third phase is *application*, wherein post-analysis data is treated as knowledge by which decisions are made.[84] A key concern in application is the re-purposing of data. Although anonymous groups are targeted, it is individuals who ultimately are affected.[85] For example, Thielman reports that personal medical information given to doctors might be anonymised before being sold, but then data-miners will commingle that information with other data sets, including public records, to create targeted advertising for pharmacies.[86] Of more prominent notoriety are the activities of marketing firms, such as the fallen Cambridge Analytica, who combined online quiz data, social media data, with polling data,

[80] Robert Madge, 'Five Loopholes in the GDPR' (MyData Journal, 27 August 2017) <https://www.dataprotection.ie/docs/Anonymisation-and-pseudonymisation/1594.htm> accessed 11 December 2017.

[81] "The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person." See *Regulation* (n 14), Recital 26.

[82] Article 4(5) defines 'pseudonymisation' as:

> the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

See ibid.

[83] Including that used for statistical or research purposes. ibid, Recital 26.

[84] Oostveen (n 52) 307–308.

[85] ibid 307.

[86] Sam Thielman, 'Your Private Medical Data is for Sale—and It's Driving a Business Worth Billions', *The Guardian* (10 January 2017) <http://www.theguardian.com/technology/2017/jan/10/medical-data-multibillion-dollar-business-report-warns> accessed 6 December 2017.

to identify behaviours linked to voting habits, allowing them to create targeted advertising during election campaigns.[87]

Automated data application poses further problems with concerns being raised over the obfuscation of accountability for discrimination and unjust outcomes as a result of AI decisions.[88] Early manifestations of automated decision-making have yielded alarming outcomes, with Microsoft's Twitter-fed AI bot, Tay, famously tweeting that "Hitler did nothing wrong", and that feminists "should all die and burn in hell".[89] It is evident that such systems are still reliant upon quantitative input from the Internet over qualitative factors. Increasingly, big data will start to supplant actual human judgment in law enforcement, judicial and healthcare scenarios. Algorithmic technology is already in use in the USA through the COMPAS tool, which has proven to perpetuate discrimination within the criminal justice system.[90] Although Article 22 of the GDPR sets down a right "not to be subject to a decision based solely on automated processing", this right clearly will not be of great consequence retroactively, where that decision has had life-altering changes or even death.

V. Informational Decolonisation

A. Informational violence[91]

Can we argue that the *application* of data is a breach of privacy laws? It is hard to accept that millions of US citizens consensually parted with their lifestyle details whilst fully understanding the ramifications in political advertising that they would later see. Nor is it likely that patients would be content for information discussed in the privacy of the doctor's surgery to result in advertising relating to those very ailments. Whilst privacy-related legal protections are wide-ranging, the commercial and political *use* of that data is still left somewhat unaccounted for. If we think of a data protection breach as the intrusive observation of an individual's

[87] Cambridge Analytica, 'Cambridge Analytica – About Us' (30 September 2015) <https://cambridgeanalytica.org/about> accessed 18 February 2017.

[88] See Bingham Centre for the Rule of Law (n 53) 3; Mady Delvaux, 'Report with Recommendations to the Commission on Civil Law Rules on Robotics' (Committee on Legal Affairs – The European Parliament 27 January 2017).

[89] Alex Hern, 'Microsoft Scrambles to Limit PR Damage over Abusive AI Bot Tay', *The Guardian* (24 March 2016) <http://www.theguardian.com/technology/2016/mar/24/microsoft-scrambles-limit-pr-damage-over-abusive-ai-bot-tay> accessed 13 July 2018.

[90] Bingham Centre for the Rule of Law (n 53) 3.

[91] Portions of this section have been adapted from an unpublished LL.M. paper by the same author entitled: 'Oscillations of Identity, Violence and Sovereignty in Cyberspace'.

personal information by unwelcome eyes, behavioural targeting is the intrusive application of intelligence based on users' personal data.

Even the notion of autonomy expressed in Recital 7 of GDPR, whereby "[n]atural persons should have control of their own personal data" relies on the oxymoron of the natural person having access to the data. Digital denaturalisation, much like its physical counterpart, constitutes a loss of autonomy. While data protection is founded in the *appropriation* of data, the *transmission* of information, post-analysis, remains unaccounted for in law.

Although the technology and method are very different to subliminal advertising, now prohibited in most jurisdictions including the EU,[92] the upshot is not altogether different: a form of informational violence based on subtle psychological techniques of persuasion with the effect of distorting fair competition.

Informational violence can manifest in various forms. As Cybenko and others note, hacking techniques—such as Denial of Service attacks—interfere with computational processes and violate human *property*.[93] By contrast, informational violence impedes upon the cognitive processes of individuals. Informational violence thus constitutes a gradual and palpitating trespass into the solitary confine of the mind.

On the web, oscillations of threat and pleasure normalise what researchers at the University of Turin call "voluntary servitude"[94] that negate each other to create unconcern towards the machineries at play. Network leviathans prosper from this indifference. Each network comprises a vast anatomy of interconnected nodes that may apprehend practically limitless quantities of informational knowledge through AI, but continue to rely on humans to develop emotional intelligence. To this end, users provide the emotional system of the artificial humanoid of the web. Gerlitz and Helmond call this a "like economy", which now encompasses a range of emotions, from laughter, anger, sadness, surprise, and so forth.[95] From consensually acquired emotional intelligence, human weaknesses may thus be repurposed through automated transformation. Although, under Article 6 of the GDPR, repurposing is now prohibited, the bewildering array of privacy options

[92] An EEC directive of 1989 prohibited subliminal advertising. See Council Directive 89/552/EEC of 3 October 1989 on the coordination of certain provisions laid down by Law, Regulation or Administrative Action in Member States concerning the pursuit of television broadcasting activities 1989, Article 10(3).

[93] George Cybenko, Annarita Giani and Paul Thompson, 'Cognitive Hacking' in Marvin Zelkowitz (ed), *Advances in Computers*, vol 60 (1st edn, Academic Press 2003) 59.

[94] Alberto Romele, Francesco Gallino, Camilla Emmenegger and Daniele Gorgone, 'Panopticism is Not Enough: Social Media as Technologies of Voluntary Servitude' (2017) 15, 208–209.

[95] Carolin Gerlitz and Anne Helmond, 'The Like Economy: Social Buttons and the Data-Intensive Web' (2013) 15 New Media & Society 1349.

on every site is likely to produce what Schwab calls "consent fatigue",[96] whereby unconscionable policies will be accepted as the path of least resistance.[97]

Armed with emotive data, marketers and technology firms, from psychographic profiles, assign labels to individuals that rigidify algorithmic social categories, which as Mejias states, has the effect of foreclosing identities from variation.[98] Identity, for Deleuze, is formed through the resemblance of variations, yet within each variation are lesser variations that contradict similarity.[99] Big data's social labels do not presently account for such nuance. The violent biopolitical history of the 20th century already provides a lesson in social labelling. Yet these designations not only externally objectivise users, but as a 2016 study revealed, they have the dangerous capacity to influence self-perceptions based on trust in algorithmic reliability.[100]

Social networks further their habit-forming effects though techniques from the gambling industry, such as the scrolling gesture that replicates the reward-or-loss tenacity of slot machines.[101] Facebook's colour scheme also employs methods of visual science, whereby blue engenders trust and dependability, while its red notifications instigate urgency, strengthening its addictive qualities.[102] The site's co-founder Sean Parker acknowledges that the occasional dopamine hit of the liked photograph or post creates a "social validation feedback loop"[103] that encourages

[96] Pierre-Nicolas Schwab, '30 Days to Read Privacy Policies: Consent Fatigue Will Make GDPR Ineffective' (*Into the Minds*, 24 May 2018) <http://www.intotheminds.com/blog/en/30-days-to-read-privacy-policies-consent-fatigue-will-make-gdpr-ineffective/> accessed 6 June 2018.

[97] ibid; Richard H Thaler and Cass R Sunstein, Nudge (Yale University Press 2008) 35.

[98] Mejias (n 63) 83.

[99] Gilles Deleuze, *Difference and Repetition* (Columbia University Press 1994) xix.

[100] Christopher A Summers, Robert W Smith and Rebecca Walker Reezek, 'An Audience of One: Behaviorally Targeted Ads as Implied Social Labels' (2016) 43 Journal of Consumer Research 156, 171.

[101] Mattha Busby, 'Social Media Copies Gambling Methods "to Create Psychological Cravings"', *The Guardian* (8 May 2018) <http://www.theguardian.com/technology/2018/may/08/social-media-copies-gambling-methods-to-create-psychological-cravings> accessed 16 May 2018.

[102] Leo Widrich, 'Why Facebook Is Blue: The Science of Colors in Marketing' (*Social*, 25 April 2015) <https://blog.bufferapp.com/the-science-of-colors-in-marketing-why-is-facebook-blue> accessed 25 May 2018.

[103] Erica Pandey, 'Sean Parker: Facebook Was Designed to Exploit Human "Vulnerability"' (*Axios*, 9 November 2017) <https://www.axios.com/sean-parker-facebook-was-designed-to-exploit-human-vulnerability-1513306782-6d18fa32-5438-4e60-af71-13d126b58e41.html> accessed 25 May 2018.

daily activity. The social network thus promises reward in tandem with personally relevant novelty to produce a sense of belonging.[104]

In the current climate, much of the effects are well publicised, yet billions of users continue to engage on social media platforms regardless. In spite of the public controversies surrounding the practices of Cambridge Analytica and its parent company, the Strategic Communication Laboratories (SCL) Group,[105] Facebook boasted record revenues of $11.97 billion in the first quarter of 2018.[106] At the time of writing, investigations are being carried out into election meddling,[107] misuse of funds, illegal data appropriation and sharing,[108] but there remains no policy discussion relating to the post-processing relay of psychologically manipulative advertising itself. Thus far, a spate of dramatic 'tech trails' have taken place within the walls of legislatures, at the US Senate,[109] the EU Parliament,[110] and a UK Select Committee,[111] where testimony takes place without oath and with privilege against defamation. Online behavioural advertising is presided over by the self-regulatory Advertising Standards Authority, a private company,[112] whose rules remain founded in the collection of data and not delivery.[113]

[104] Björn Enzi, Moritz de Greek, Ulrike Prösch, Claus Tempelmann and Georg Northoff 'Is Our Self Nothing but Reward? Neuronal Overlap and Distinction between Reward and Personal Relevance and Its Relation to Human Personality' (2009) 4 PLoS ONE, 7.

[105] The Guardian, 'The Cambridge Analytica Files', *The Guardian* (17 March 2018) <https://www.theguardian.com/news/series/cambridge-analytica-files> accessed 8 June 2018.

[106] Facebook, Inc., 'Facebook Reports First Quarter 2018 Results' (25 April 2018).

[107] Jeremy White, 'Federal Trade Commission "Investigating Facebook after Cambridge Analytica Scandal"', *The Independent* (21 March 2018) <https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-cambridge-analytica-federal-trade-commission-ftc-investigation-privacy-rules-consent-decree-a8265906.html> accessed 8 June 2018.

[108] ICO, 'ICO Statement: Investigation into Data Analytics for Political Purposes' (*Information Commissioner's Office*, 3 May 2018) <https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2018/05/ico-statement-investigation-into-data-analytics-for-political-purposes/> accessed 8 June 2018.

[109] Bloomberg Government (n 59).

[110] Jennifer Rankin, 'Complaints that Zuckerberg "avoided Questions" at European Parliament', *The Guardian* (22 May 2018) <http://www.theguardian.com/technology/2018/may/22/no-repeat-of-data-scandal-vows-mark-zuckerberg-in-brussels-facebook> accessed 8 June 2018.

[111] Commons Select Committee, 'Alexander Nix to Appear Again before the Committee' (*UK Parliament*, 7 June 2018) <https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/fake-news-nix-evidence-17-192/> accessed 8 June 2018.

[112] Advertising Standards Authority, 'About ASA and CAP' (*Advertising Standards Authority*, 2 March 2017) <http://www.asa.org.uk/about-asa-and-cap.html> accessed 8 June 2018.

[113] Advertising Standards Authority, 'Appendix 3 Online Behavioural Advertising' (*Advertising Standards Authority*, 21 November 2012) <http://www.asa.org.uk/type/non_broadcast/code_section/appendix-3.html> accessed 8 June 2018.

B. Informational self-determination

How then can human rights law respond to the informational violence of data technologies? And where lies the threshold of tolerance between innocuous advertising and psychological interference? In 1982, the German *Bundestag* legislated a population census that sparked controversy out of public fears that personal information could later be repurposed.[114] A year later, a class-action challenge was launched at the *Bundesverfassungsgericht* (the German Constitutional Court)[115] that lead to the creation of a principle called the 'right to informational self-determination to distinguish this problem from that of privacy.[116] As Rouvroy and Poullet stress, the right should not be mistaken for the right to maintain autonomy over one's own information,[117] a privacy issue now addressed by GDPR. Rather, the German court conceived informational self-determination as being the control of one's data as a means to ensure an autonomous existence as a citizen.[118] Hornung and Schnabel note that privacy and informational self-determination are interdependent but ultimately distinct legal matters.[119] In terms of the rights' application to modern data catalysis, it would be useful to prevent private entities and political parties from manipulating individuals based on their psychological weaknesses.

Although the freedoms of thought and conscience in Article 9 of the ECHR and Article 18 of the International Covenant on Civil and Political Rights (ICCPR) customarily are applied to religious issues, they also function to protect the manifestation of personal, political, philosophical, and moral beliefs,[120] a logical interlacement given that secular convictions neurologically occur in the same part of the brain as religious beliefs.[121] In the political context, a right to informational

[114] Gerrit Hornung and Christoph Schnabel, 'Data Protection in Germany I: The Population Census Decision and the Right to Informational Self-Determination' (2009) 25 Computer Law & Security Review, 85–87.

[115] Volkszählungsurteil BVerfGE 65,1 Bundesverfassungsgericht, 15 December 1983, 1 BvR 209, 269, 362, 420, 440, 484/83.

[116] Hornung and Schnabel (n 114) 85–86.

[117] Antoinette Rouvroy and Yves Poullet, 'The Right to Informational Self-Determination and the Value of Self-Development: Reassessing the Importance of Privacy for Democracy' in Serge Gutwirth et al (eds), *Reinventing Data Protection?* (Springer Netherlands 2009) 51.

[118] ibid 45–46; Hornung and Schnabel (n 114) 86.

[119] Hornung and Schnabel (n 114) 86.

[120] Jean-François Renucci, 'Article 9 of the European Convention On Human Rights: Freedom of Thought, Conscience and Religion' (Human Rights Files, Council of Europe 2005) 12–13.

[121] Neuroscientists from UCLA have proven that belief in religious and secular ideas occur in the ventromedial prefrontal cortex. See Allison Bond, 'Belief in the Brain' (*Scientific American*, 1 March 2010) <https://www.scientificamerican.com/article/belief-in-the-brain/> accessed 13 December 2017.

self-determination would constitute a hybridisation of Article 3 of the ECHR's Protocol, which requires that elections be held "under conditions which will ensure the free expression of the opinion of the people in the choice of the legislature" with Article 9. The recent examples of psychological manipulation by shadowy forces during elections[122] will serve as a lesson learned from these early days of big data.

V. Relocating the Spatiality of Data Rights

The potential uses and misuses of data catalysis to human well-being and democracy are, at this early stage, as diverse as our own imaginations, and we can only guess where it will all go. How then might human rights be refurbished to respond reflexively to the overreaching of present and future technological phenomena?

The first area of development would be the enlargement of the right to data protection and associated rights, from being fundamental rights, to universal human rights. To account for the problem of territorial scope, the GDPR addresses jurisdiction comprehensively by expanding its reach to processors both in and outside of EU.[123] The problem, of course, is the practical reality of rogue organisations complying with these rules, and the enforceability of international law in third countries that do not provide adequate data protection. Therefore the direct and horizontal effect of fundamental rights of a EU Regulation is a considerable step forward for data subjects within the Union. Yet beyond terrain of the physical, data's ethereality becomes more apparent, revealing the material limits of state sovereignty.

Data's true terrain is, of course, cyberspace. Fletcher observes the disjuncture between the real and cyber as being based upon arbitrarily constructed yokes between the individual and socio-cultural artefacts.[124] If we are to dissolve such linkages and assess their empirical influence on social reality, their effects only manifest within the material and biological real. Wertheim draws a practical unification between the real and cyber by observing that the imaginary of

[122] Carole Cadwalladr, 'The Great British Brexit Robbery: How Our Democracy Was Hijacked', *The Observer* (7 May 2017) <http://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy> accessed 13 December 2017.

[123] Data protection applies where goods and services are offered to EU citizens, or where non-EU organisations monitoring data subjects within the EU. It also applies to processors that are established both in the EU and outside, irrespective of whether the processing occurs in the jurisdiction or not. Data transfers outside of the union must be to a third country deemed to offer adequate protection by the commission.

[124] Gordon Fletcher, 'Between Heaven and Charing Cross? Cyberspace as Urban Space' (*Spaceless*, 1995) <http://www.spaceless.com/papers/20.htm> accessed 20 November 2017.

cyberspace merely reiterates the Abrahamic reification of the celestial space as a foundation for law.[125] In the mediaeval imaginary space, there was regulation of the body *and* the soul, she writes. A critical factor at the time was that the universe was deemed finite, and could be quantified, albeit arbitrarily. Beneath the celestial strata of St Thomas Aquinas' jurisprudential taxonomy, lay the behavioural traits of all rational creatures that he ascribed as the natural law. With the tacit mandate of an omniscient Creator, these universal laws furnished the church and rulers with the building blocks of their own positive laws.

In the digital imaginary of the Internet, a different form of all-knowing entity exists in a new kind of cloud. Here, there is no jurisdiction without body. Much like Agamben's notion of the State of Exception as a legally sanctioned zone of lawlessness,[126] data catalyses are able to strip the user of political identity and autonomy—her *bios*. The user, then, is kettled into subgroups on the Internet whereupon they are bestialised into a commodified form of digital *zoē*,[127] where terrestrial laws are ineffectual. This ever-proliferating paradigm has snowballed in the name of the business efficacy. Algorithms, as Galloway insists, are firmly monolithic in their advancement of sanitised institutionalism,[128] yet by unlocking the secrets to human nature, data processes may be rationalised under the premise of collective betterment. Big data is the new natural law.

In practical terms, this suggests that laws on data need to be re-conceived as if cyberspace were real space. In 2016, the United Nations Human Rights Council passed a Resolution for "promotion, protection, and enjoyment of human rights on the Internet" emphasising "that the same rights that people have offline must also be protected online, in particular freedom of expression, which is applicable regardless of frontiers".[129] Although a positive development, the Resolution was not binding and saw 14 countries vote against with 13 abstentions.[130] Clearly, for cyber law to be effective, the option for states to abstain creates a void where rules may be broken. As d'Amato states, the notion of consent in international law flies in the face of its purpose, as it gives states the prerogative not to adopt and ratify

[125] Margaret Wertheim, *The Pearly Gates of Cyberspace: A History of Space from Dante to the Internet* (1st ed, WW Norton 1999) 18–43.

[126] Giorgio Agamben, 'The State of Exception as a Paradigm Of Government' in *State of Exception* (University of Chicago Press 2005) 1–31.

[127] Giorgio Agamben, 'The Politicalization of Life' in *Homo Sacer: Sovereign Power and Bare Life* (Stanford University Press 1998) 9–14.

[128] Alexander R Galloway, *The Interface Effect* (Polity 2012) 99.

[129] The Promotion, Protection and Enjoyment of Human Rights on the Internet A/HRC/32/L20 (United Nations 2016).

[130] Maëli Astruc, 'UN Human Rights Council Takes Actions On Internet Rights, Corporations' (*Intellectual Property Watch*, 14 July 2014) <https://www.ip-watch.org/2014/07/14/un-human-rights-council-adopts-resolutions-on-internet-corporate-responsibility/> accessed 15 December 2017.

legal principles that they find inconvenient.[131] Given the global impact of the Internet, localised norms do not suffice, allowing for private data firms and their funders to amass dangerous amounts of power around the world.[132]

The right to self-determination is one of the most ubiquitous human rights principles of all. Article 1 in both the ICCPR and the International Covenant on Economic, Social and Cultural Rights (ICESCR) state: "All peoples have the right of self-determination. By virtue of that right they freely determine their political status and freely pursue their economic and cultural development". After World War II, the right emerged as the primary tool to instigate decolonisation.[133] It since has been recognised by the international community as a peremptory norm binding upon all states.[134] Read outside the decolonialisation context, the words of Article 1 can be transposed to embrace the freedom to maintain one's political autonomy without interference, and freedom of economic and cultural choice. Informational self-determination as a universal peremptory norm, hence, will set a global standard for an even more politically, culturally and economically well-playing field, all of which are, of course, interlinked. It is clearly unrealistic to impose an outright ban on targeted advertising. A suitable threshold would therefore be to apply the principle where the thing advertised concerns the public at large.

Data processing is just one cogwheel in the techno-leviathan that is in dire need of regulation. The monoliths of the skyscraper and the black smartphone emerged as seemingly indestructible cultural institutions of the post-War capitalist era. Yet the manner in which both have been attacked signifies an unforeseen defencelessness within the neoliberal machine wherein any party with resources has been able to meddle.

The non-territorial ethereality of data is ample foundation for establishing international checks and balances, such as a World Court of Human Rights (as suggested by Nowak[135]) and an international convention that imposes binding

[131] Anthony d'Amato, 'Is International Law Really Law?' (1984) 79 Northwestern University Law Review 1293, 1309.

[132] Jack Lewis, 'Cambridge Analytica Endangers Global Democracy, and It Must Be Stopped', *The Diamondback* (20 November 2017) <http://www.dbknews.com/2017/11/21/cambridge-analytica-endangers-global-democracy-and-it-must-be-stopped/> accessed 14 December 2017.

[133] Although presented under the pretext of democracy, the post-War decolonial drive was founded by the US desire to break up the European stronghold of the south during the Cold War and to strengthen pro-capitalist numbers at the UN. See Leslie James and Elisabeth Leake (eds), *Decolonization and the Cold War: Negotiating Independence* (New approaches to International History, Bloomsbury 2015) 1–2.

[134] Responsibility of States for Internationally Wrongful Acts A/56/49(Vol I)/Corr4 (United Nations 2001).

[135] Manfred Nowak, 'The Need for a World Court of Human Rights' (2007) 7 Human Rights Law Review, 251–259.

human rights obligations on private corporations (as proposed in the Lima Declaration[136]). All such propositions are objects of a study unto their own, but notwithstanding other innumerable imperatives for such normative legal developments, data law itself is a pervading argument in their favour.

VI. Conclusion

In evolutionary terms, the Internet is the Earth's youngest wilderness. To contemporary powers it is a virgin territory where those who control its quarries may flourish under their own stipulations. Whilst the wanton *excavation* of its resources is seeing a steady sharpening of supervision by international common sense, the *exploitation* of those resources remains a prerogative in which highest bidders may luxuriate. Through efficient tools of mechanised processing, 'stock' is intermingled and bestialised through collective attributes, beyond law's grasp. In this indeterminate zone of *terra nullius*, individual personhood, in a denaturalised state of spectrality, is put to work in data's industrial complex, alienating the final material production from the natural human resources whence it came.[137] The final output, then, is used to automate new forms of natural law employed to further subjugate the very same populace.

Much like the civilising missions in the Age of Discovery, the maladies of data's informational violence will only be felt long after being diagnosed. The mutually beneficial privity of contract between data processors ensures that data's unseen processes evade human rights jurisdiction. International instruments and domestic law, hence, must reimagine cyberspace as the proximity of real territory, much like the seas and air, so as to prevent unprecedented abuses from these juridical voids. As Fanon wrote, during the process of decolonisation, the indigenous population were "discerned only as an indistinct mass".[138] We all are the indigenous population of cyberspace. A process of decolonisation has begun, but it still has much of a way to go before the mass once again are recognised as constituents of reality.

[136] Worldwide Movement for Human Rights (FIDH, 'Lima Declaration on Human Rights and Business' (2012) 1–6.

[137] Of alienation, Karl Marx wrote:

> The alienation of the worker in his product means not only that his labour becomes an object, an external existence, but that it exists *outside him*, independently, as something alien to him, and that it becomes a power on its own confronting him; it means that the life which he has conferred on the object confronts him as something hostile and alien.

See Karl Marx, *Economic and Philosophic Manuscripts of 1844* (Dover Publications 2007) 67–83.

[138] Frantz Fanon, *The Wretched of the Earth* (Constance Farrington tr, Penguin Books 2001) 34.

MNCs and the Human Rights Regulatory Challenge: A Critique of 'Integrated Theory of Regulation' and the Case for a Possible Alternative

Samuel E. Ojogbo*

I. Introduction

As the major drivers of globalisation, multinational corporations (MNCs)[1] are in large part responsible for the benefits associated with it. The International Chamber of Commerce has suggested that "globalisation has made the world economy more efficient and has created hundreds of millions of jobs, mainly, but not only, in developing countries".[2] This does not mean that globalisation and the activities of MNCs are all about global economic progress and job creation in developing markets, because the activities of MNCs also produce certain adverse

* Lecturer in Law, Benson Idahosa University. Ph.D. (Nottingham), LL.M. (Western), LL.B. (LASU).

1 The acronym MNCs will be used to represent "multinational corporations", the preferred term used throughout this article to denote major multinational business corporations. Other terminologies such as "multinational enterprises" (MNEs) and "transnational corporations" are also used to identify multinational (global) corporate business operations, but the differences in terminologies is not material to discussions on the impact of global corporate business activities in developing markets. For a definition of the various terminologies and their origin, see Peter M Muchlinski, *Multinational Enterprises and the Law* (2nd ed, OUP 2007) 5–8.

2 International Chamber of Commerce, 'Brief on Globalization' (Organization for Security and Co-operation in Europe, November 22, 2000) <https://www.osce.org/secretariat/42286?download=true> accessed 5 December 2017.

effects that impact upon the human rights of communities when they operate in developing markets.[3]

It is this adverse effect of MNC activities, and the lack of adequate remedies for the victims, that are the focus of this article. This has become especially relevant given the frequency with which non-governmental organisations (NGOs) and activists have publicised instances of business conduct that has violated universally agreed upon human rights norms since the last decade of the last century.[4] The outrage that followed some of the major human rights and environmental disasters in developing markets[5] is one of the major factors that has influenced the development of new initiatives on Business and Human Rights (BHR), which has led to the emergence of various Corporate Social Responsibility (CSR) codes by some major MNCs and Voluntary Codes of Conduct (VCC) by inter-governmental institutions since the 1990s.[6]

However, notwithstanding the growing body of regulatory initiatives that seek to make MNCs accountable, environmental damage and human rights breaches

[3] Muchlinski (n 1) 487–489: "When operating in developing countries, where comparable employers may not exist, MNEs should provide the 'best possible wages, benefits and conditions of work, within the framework of government policies'. These should be related to the 'economic position of the enterprise, but should be at least adequate to satisfy basic needs of the workers and their families'…"

[4] See, for example, *Human Rights Watch and Center for Human Rights & Global Justice*, 'On the Margins of Profit: Rights at Risk in the Global Economy' (Human Rights Watch, February 2008) <http://www.hrw.org/sites/default/files/reports/bhr0208webwcover.pdf> accessed 19 October 2017; *Human Rights Watch*, 'The Price of Oil: Corporate Responsibility and Human Rights Violation in Nigeria's Oil Producing Communities' (Human Rights Watch, January 1999) <http://pantheon.hrw.org/reports/1999/nigeria/nigeria0199.pdf> accessed 19 March 2016.

[5] Some examples of environmental and human rights disasters that have generated global outrage include the destruction of Ecuadorian Amazon by Texaco: see *The Inter-American Commission on Human Rights*, 'Report on the Situation of Human Rights in Ecuador' reprinted in Basic Documents Pertaining to Human Rights in the Inter-American System OEA/Ser L V/II.96 Doc 10 Rev 1 (1997); the activities of Shell that caused the environmental degradation in Ogoni and the eventual hanging of Ken Saro Wiwa and his eight Ogoni kinsmen on 1 November 1995: see '1995: Nigeria Hangs Human Rights Activist' *BBC News* (London, 10 November 1995) <http://news.bbc.co.uk/onthisday/hi/dates/stories/november/10/newsid_2539000/2539561.stm> accessed 26 January 2015; and the collapse of a factory building in Sava, Bangladesh, 13 March 2013: see 'Bangladesh building collapse death toll passes 500' *BBC News* (London, 3 May 2013) <http://www.bbc.co.uk/news/world-asia-22394094> accessed 2 March 2016.

[6] Christen Broecker, '"Better the Devil You Know": Home State Approaches to Transnational Corporate Accountability' (2008–2009) 41 NYU J. Int'l L & Pol 159, 160.

are still prevalent in developing markets.[7] This has led to the conclusion by some commentators that existing legal and regulatory regimes for regulating MNCs are inadequate.[8] As a result, commentators have increasingly focused on the obligation of MNCs under existing international norms, especially human rights law, as a means of regulating MNCs to make them accountable for their human rights violations.[9] Some scholars have developed alternative regulatory theories in this regard. One of them is the main subject of this investigation: the "integrated theory of regulation" developed by Deva Surya to solve what he termed the difficulty in regulating a difficult regulatory target—MNCs.[10]

Deva and most of the current human rights law scholarship are agreed on the need for a legal or regulatory framework that could address the current situation of corporate impunity for human rights violations in developing markets. The complex nature of modern MNCs and their global operations make them a difficult regulatory target because they operate in developing markets through their subsidiaries, which often take the form of special purpose vehicles (SPVs). However, it is argued that the law that regulates their entry into a developing market could also be relied upon to regulate their conduct within the jurisdiction with a view to preventing corporate human rights violations. The place of domestic law, especially corporate law,[11] as a regime for addressing corporate responsibility in this regard will be discussed later on, but in order to be able to effectively challenge

[7] See the *UN Commission on Human Rights* (UNCHR), 'Interim Report of the Special Representative to the Secretary General on the Issue of Human Rights and Transnational Corporations and other Business Enterprises' (22 February 2006) E/CN.4/2006/97 para 30; see also: United Nations Environmental Programme (UNEP), 'Environment Assessment of Ogoniland (UNEP, 2011) <https://www.zaragoza.es/contenidos/medioambiente/onu//issue06/1130-eng.pdf> accessed 23 January 2015.

[8] Sukanya Pillay, 'And Justice for All? Globalization, Multinational Corporations, and the Need for Legally Enforceable Human Rights Protection' (2003–2004) 81 U Det Mersy L Rev 489, 522; Surya Deva, *Regulating Corporate Human Rights Violations: Humanising Business* (Routledge 2012) 12.

[9] See, for example, Michael K Addo, 'Human Rights and Transnational Corporations—An Introduction' in Michael K Addo (ed), Human Rights Standards and the Responsibility of Transnational Corporations (The Hague: Kluwer Law International 1999); Chris Jochnick, 'Confronting the Impunity of Non-State Actors: New Fields for the Promotion of Human Rights' (1999) 21 *Human Rights Quarterly* 56; Pillay (ibid); Deva (ibid).

[10] Deva (n 8) 50–51.

[11] Corporations and companies will be used interchangeably throughout this article to represent the for-profit business structure that provides its "owners" (members/shareholders) with limited liability. Corporate law and company law will also be used interchangeably to represent the law that regulate them. We acknowledge that in certain jurisdictions, notably the US, different legal regimes regulate companies and corporations but such difference is not material to the discussion here because our reference jurisdictions in this article—the UK and some major developing markets that operate as common law jurisdictions, such as India, South Africa and Nigeria—do not operate different corporate law regimes for companies and corporations. For an analysis of the differences, see Joseph Shade, Business Associations in a Nutshell (3rd edn, West Academic Publishing 2010) 46–50.

Deva's views on the subject of regulating MNCs, it is important to first identify and analyse the basis for his views.

Deva identifies three sets of question that represent what he terms the three broad challenges to the goal of humanising business—why, what and how (the WWH challenge).[12] First, why corporations should have human rights responsibility; secondly, what the exact nature and scope of corporate human rights responsibility is; and thirdly, how corporations, especially MNCs, can be held accountable for human rights violations.[13] To resolve these challenges, he begins by identifying the existing problems. He employs Bhopal as a case study to investigate how MNCs are able to violate human rights and escape liability in developing markets by exploiting the loopholes in existing regulations.[14]

Deva's choice of Bhopal, a remote Indian town and the site of a disastrous toxic gas leakage during the night of 2 December 1984 from the methyl isocyanate storage tank of the Bhopal chemical plant owned and operated by Union Carbide India Limited (UCIL), a subsidiary of Union Carbide Company (UCC),[15] is important to the issues addressed in this article for two reasons. First, it reveals the manifest failure of traditional corporate law concepts to deal adequately with the MNC phenomenon.[16] Secondly, it brings to the fore some of the factors that make MNCs operating in developing markets a difficult regulatory target.

Deva extensively discusses how the control of UCIL by the parent UCC contributed to the accident that occurred in Bhopal.[17] He is inconclusive as to whether the manner in which UCIL, as the Indian subsidiary, was controlled played an important part in that accident, or that a different corporate governance structure would probably have produced a different outcome. Thus, the overall aim of this article is to identify how the interaction between the corporate governance structure and the actual manner of control can be a basis for suggesting an alternative approach to integrated regulation that could promote more responsible corporate behaviour, and thereby minimise corporate human rights abuses in developing markets.

The rest of this interrogation is spread across four main parts. Part II will discuss Deva's case study, Bhopal. This will provide a rational basis to analyse the

[12] Deva (n 8) 1.

[13] ibid.

[14] ibid 2.

[15] Union Carbide Company (UCC) is the American parent company of Union Carbide India Limited (UCIL). The acronyms UCC and UCIL will be used throughout this article to represent the Union Carbide Company and Union Carbide India Limited. See Deva (n 8) 29.

[16] Muchlinski (n 1) 321.

[17] Bhopal represents the gas leak in the Union Carbide Corporation (UCC) plant in Bhopal India in 1984 in which over 15,000 (fifteen thousand people) died. Deva (n 8) 24–45.

challenges to human rights regulation in developing markets. Part III will review the three levels of the theory while Part IV will critique the theory and identify the basis for an alternative approach to integration. Part V concludes the article.

II. The Bhopal Case Study as the Basis for Integrated Theory of Regulation

The power and economic influence of UCC and the demand for foreign investment played a major part in the post-entry negotiations between UCC and its host (India), as well as the manner in which the company operated in the country. According to Deva, the company was able to capitalise on the Indian Government's desire to industrialise by exploiting existing laws and effectively bypassing their application of it to its own operation in a spectacular fashion. First, the Industrial Development and Regulation Act 1951 reserved the formulation of pesticide activities for small local Indian firms, and the Foreign Exchange Regulation Act 1973 also limited foreign ownership of Indian firms to 40%.[18] UCC was able to bypass the obstacles imposed by both pieces of legislation to establish UCIL, a company engaged in activities reserved for local firms. It is important to understand the effect of majority control in corporate decision-making in order to fully appreciate why it was important to UCC and its contribution to the Bhopal disaster.

It is a general rule of company law that ownership of shares in a company having share capital qualifies the holder as a member of the company.[19] In other words, shareholding is synonymous with membership.[20] The desire for majority control by UCC in this case is instructive. Company law virtually separates ownership and control and the key players in the formal decision-making structure of companies, especially public companies, are a group called the 'board of directors'.[21] Thus, the powers of the shareholders to initiate corporate actions is very limited, as they are only entitled to approve or disapprove a few board actions.[22]

It is noteworthy that the separation of ownership and control is partial or non-existent where there is majority control that is where the corporation has a

[18] ibid 26.

[19] See *Companies Act 2006* (hereinafter, "UK CA 2006"), s 112; Companies & Allied Matters Act (Cap C20) (Laws of the Federation of Nigeria 2004) (CAMA), s 79; The Companies Act 2013 (No 18 of 2013) (hereinafter, "Indian CA"), ss 2(55) and 45–50.

[20] Derek French *et al*, *Mayson, French and Ryan on Company Law* (34th edn, OUP 2013) 165.

[21] *Aranson v Lewis* 473 A.2d 805 (Del 1984) 811; UK CA 2006 (n 19), s 71; CAMA (n 19), ss 244 and 63(3); Stephen Bainbridge, *The New Corporate Governance in Theory and Practice* (OUP 2008) 4.

[22] Bainbridge (ibid).

dominant shareholder who owns more than 50% of the outstanding voting shares.[23] In this case a shareholder is able to control the board and the management of the company. This is the typical structure of a corporate group involving a MNC and its group members (subsidiary or subsidiaries), usually structured in a 'pyramid' hierarchical form of ownership, where a parent company wholly owns (or holds the majority shares in) the subsidiary or subsidiaries and may in fact dominate their management.[24]

The total control that UCC exercised over the Indian subsidiary UCIL, which was achieved by persuading the Indian government to grant them an exemption from the 40% rule,[25] merits further discussion because of the implication of foreign control to the Bhopal disaster. UCC did in fact exercise real control over the subsidiary, UCIL, a company engaged in carrying on a hazardous activity at the Bhopal plant. Thus, based on its ownership of the majority of equity (50.9%) in the subsidiary,[26] UCC controlled the composition of the board of directors of UCIL and also had full control over its management.[27] The control by the parent company extended beyond representation on the board of directors "to the taking of key decisions regarding issues such as technology, plant design, safety, storage and handling of MIC,[28] training of employees and financial viability of the firm".[29]

This excessive "centralisation not only resulted in a rift between the formulation of 'global' policies and their 'local' implementation, but also contributed to a communication and management gap between UCC and the and UCIL".[30] J. Cassels explains how this rift played a part in the occurrence of Bhopal thus: "[s]afety information was not properly communicated from the head office, and what information was communicated was ignored".[31] In addition, there was also the problem of differences in the expectations of the parent MNC and the Indian government, because the Indian subsidiary did not meet the profit expectation of the parent company.[32] As a result, UCC was not very interested in the proper management or successful running of its subsidiary UCIL and thus resorted to cost-cutting measures. The company adopted inferior standards in terms of its

[23] Adolf A Berle and Gardiner C Means, *The Modern Corporation & Private Property* (10th reprint, Transaction Publishers 2009) 84–116.
[24] Irit Mevorach, *Insolvency within Multinational Enterprise Groups* (OUP 2009) 16.
[25] Deva (n 8) 26.
[26] Muchlinski (n 1) 315.
[27] ibid.
[28] MIC is the acronym for methyl isocyanate, the chemical that caused the explosion in the Bhopal plant.
[29] Deva (n 8) 28.
[30] ibid.
[31] J. Cassels, *The Uncertain Promise of Law: Lessons from Bhopal* (University of Toronto Press 1993) 20.
[32] Deva (n 8) 28.

staff training and maintenance of the Bhopal plant in comparison to the standard operated in West Virginia and other plants belonging to the parent UCC.[33]

The Indian government on the other hand preferred that the Bhopal plant should continue to operate safely but could not rein in the company, not only because it manufactured pesticides locally, but also because it provided much needed employment. At this point, issues of safety and the environment took a back seat because these were not the priorities of UCC[34] at the time. The application of inferior technology in the Bhopal plant, in addition to compromised safety standards and poor staff training, at a company like UCIL that was involved in dealing with or storing MIC, an ultra-hazardous and dangerous substance,[35] meant that Bhopal was a disaster waiting to happen.

The major problem in this unfolding situation at the Bhopal plant is that, at the decision-making level inside UCIL, there was no representative committed to defending the rights and interests of the vulnerable groups that is those exposed to the dangers of an accident, especially the workers and the local community. This is because the Indian Companies Act 2013 (hereinafter, "Indian CA"),[36] just like equivalent legislation in most developing markets, —such as Nigeria's Companies and Allied Matters Act 2004 (CAMA)[37]—and even in major economic jurisdictions (especially in the common law world),[38] recognises only two decision-making elements: the board of directors, and the company in the annual general meeting. This makes governance of the corporation fundamental to our criticism of Deva's suggested approach, and relevant to any proposal that seeks to address corporate human rights violations. Thus, it is important to mention here that decision-making inside the corporation does not generally include those most affected by the activities of the corporation, such as local communities. In the case of UCC

[33] Amnesty International, *Clouds of Injustice: Bhopal Disaster 20 Years On* (Amnesty International 2004) 42–43.

[34] Deva (n 8) 29.

[35] UCC's *Reactive and Harzadous Chemicals Manual*, states that MIC is 'a harzardous material by all means of contact' and recognised poison by inhalation', UCC, *Bhopal Methyl Isocyanate incident Investigation Team Report*, Danbury, March 1985 as quoted in Amnesty International (n 33) 11.

[36] Indian CA, ss 88–122 and 279. This is in pari materia with ss 165–192, 285–292, Companies Act 1956 (Act No 1 of 1956) (hereinafter, "Indian CA 1956"), the extant corporate legal regime in India at the time of the Bhopal incident.

[37] See CAMA, ss 211–244, 279.

[38] In the UK, the power to govern the company is shared between the shareholders and the board of directors and it is a contractual relationship and it is the articles that determines the extent of management power conferred on the board. See, UK CA 2006 (n 19), s 257; the Companies (Model Articles) Regulations 2008, SI 2008/3229, reg 2, Sch 1 art 3 (Model Articles for Private Companies Limited by Shares); reg 4, Sch 3 art 3 (Model Articles for Public Companies).

and its subsidiary, UCIL, and consistent with how many MNCs are organised and structured, the parent UCC controlled the functioning of its subsidiary.[39]

Another major factor that arguably encourages corporate irresponsibility in developing markets is the lack of access to legal remedies as a result of the inefficiency of their judicial infrastructure. In the case of Bhopal, the convoluted battle to secure compensation for the victims revealed the profound deficiency in the Indian judicial infrastructure, which is the general trend in many developing markets.[40] In the Bhopal disaster case, the Indian government, which secured the exclusive right (through the Bhopal Gas Leak Disaster (Processing of Claims) Act 1985)[41] to represent the victims, considered it best to sue the parent company UCC in a US court. UCC on the other hand pushed for the trial to take place in an Indian court where they could rely on the weakness of the system to manipulate the process in order to induce a settlement.[42] The other reason that the Indian government preferred to pursue the matter in a US court was because of the "general incompetence of the Indian courts and the legal system to handle effectively a case of this magnitude".[43]

The case which the Indian government commenced in the US court was, however, eventually dismissed on the grounds of *forum non conveniens*, as the judge declared that he was "firmly convinced that the Indian legal system is in a far better position than the American courts to determine the cause of the tragic event and thereby fix liabilities".[44] As will be illustrated, the later litigation that took place in the Indian courts confirmed not only the general incompetence earlier pleaded by the Indian government in the US court, but also the dilemma that dealing with MNCs poses for developing markets.

Two developments in the trial that took place in India highlights the sensitivities of the Indian authorities and the inefficiency of the Indian judicial infrastructure. First, the case was not vigorously prosecuted because the Government of India neither wanted its role in the Bhopal disaster exposed and subjected to international

[39] Deva (n 8) 28.

[40] An example is in Nigeria where a case of *Shell Petroleum Development Company Nigeria Ltd v Joel Anaro & Ors* (2015) LPELR-24750 (SC) concerning human rights abuse took 32 years to resolve. See Ade Adesomoju, 'Oil Spills: 32 Years After, Supreme Court Orders Shell to pay N30m Compensation' *The Punch* (Nigeria, 6 June 2015) <https://punchng.com/news/oil-spill-32-years-after-scourt-orders-shell-to-pay-n30m-compensation/> accessed 3 March 2016.

[41] Bhopal Gas Leak Disaster (Processing of Claims) Act 1985 (No 21 of 1985).

[42] Deva (n 8) 40.

[43] ibid 38.

[44] *In Re: Union Carbide Corporation Gas Plant Disaster at Bhopal, India in December 1984* (1986) 634 F Supp 842, 866.

scrutiny nor did it want the Bhopal litigation to discourage other MNCs from investing in India.[45]

Secondly, UCC was able to exploit the inefficiency of the Indian judicial infrastructure to such effect that throughout the period of litigation, from the district court through to the High Court and Supreme Court, the matter never proceeded to an assessment on the merits of the case. This is because, in the words of Deva, "UCC played its cards skilfully (*e.g.* delaying proceedings, increasing the complexity of the case, …, filing cross-appeals, challenging powers and jurisdiction of the Indian courts, and even conveying a veiled threat about the non-enforceability of an Indian judgement against UCC in the US)".[46] This is with a view to coercing "the government to enter into a settlement".[47] Thus, after years of frustrating litigation, the India Supreme Court finally approved a settlement between UCC and the Indian government by its two orders dated 14 and 15 February 1989,[48] in the following words: "[t]he aforesaid payments [US$470 million] shall be made to the Union of India as claimant and for the benefit of all victims of the Bhopal disaster… and not as fines, penalties, or punitive damage".[49]

The Bhopal case represents a typical scenario of a human rights violation in developing markets[50] and the type of redress available to the victims of such violations. In such situations, Deva argues that existing regulatory mechanisms are inadequate to deal with human rights violations and he proposes his "integrated theory of regulation" as a process of integrating different available levels of regulation to effectively deal with corporate human rights abuses.

III. Deva's Integrated Theory of Regulation

Deva identifies three levels of regulation for his integrated approach that is aimed at adequately regulating MNCs to ensure human rights integrity in developing markets. It is instructive that the underlying reason why Deva suggests an integrated approach is that the structure and operation of MNCs make them a difficult regulatory target, which cannot be adequately regulated by a single regulatory theory or strategy. However, as will be discussed later on,[51] Deva did not specifically address the important issue of how the very structure of the modern corporation may itself be fundamental to the Bhopal case and generally contribute

[45] Deva (n 8) 39.
[46] ibid 40
[47] ibid.
[48] *Union Carbide Corporation v Union of India* AIR 1990 SC 273.
[49] Deva (n 8) 275.
[50] ibid 24.
[51] See Part III.B below.

to corporate disregard for human rights; instead, he focuses on transnational regulation as a strategy for providing adequate remedy to human rights victims.

As this article proceeds to review the levels of regulation proposed by Deva, it is important to note that employing a mix of regulatory regimes is in principle a useful strategy for the effective regulation of modern multi-layered corporations operating multi-nationally. Therefore, there is no attempt here to contest the efficacy of Deva's multi-layered regulatory approach. However, the problem is that Deva's remedial focus fails to address the problem posed by the structure of the modern corporation. This is the basis for this review that aims to explore how a change in corporate governance structure may be an alternative basis for promoting international human rights standards.

A. A review of Deva's proposed regulation at the institutional level

Deva argues that there is a need for regulation at the institutional level. According to him, the place of the corporation as an important institution in society makes regulation at the institutional level imperative.[52] By regulation at the institutional level, Deva means the putting in place of codes of conduct, guidelines, principles, charters or policy statements by each business enterprise.[53] Deva distinguishes his suggested corporate code from the "self-regulating" corporate codes voluntarily adopted by most major MNCs. According to him, the proposed code is a modified version of self-regulation.[54] He identifies two aspects of modification. First, in designing the proposed code, the concerned institution is to be guided by the content of regulatory initiatives at the national and international levels, as well as by input from its stakeholders. Secondly, the stakeholders of the institution concerned will try to ensure that—through a range of strategies and sanctions—the initiatives adopted are implemented in their letter and spirit by the concerned corporation.[55]

Deva rightly anticipates that opinions may sometimes diverge between a particular MNC and its stakeholders about the contents of the code. To resolve such differences when they arise, he proposes that the institution concerned should adopt the view reflective of its position and issue an explanatory note along with the adopted code to explain the circumstances that required deviations *vis-a-vis* existing national or international regulations, or why certain suggestions of

[52] Deva (n 8) 204.
[53] ibid 204.
[54] Deva (n 8) 204, citing Michael Clark, *Regulation: The Social Control of Business between Law and Politics* (Macmillan Press Ltd 2000) 3.
[55] Deva (ibid).

stakeholders have not been included in the code.[56] The implication of resolving the divergence of opinions between a MNC and its stakeholders in this manner is that the proposed code does not necessarily have to strictly conform to any national or international regulations to guide the operations of the MNC concerned with respect to human rights, as long as the non-conformity can be explained.

There are many problems with the corporate code suggested by Deva as a regulatory tool intended to make MNCs take their human rights responsibilities seriously. First, his attempt to differentiate his own suggested code from the other corporate codes currently adopted and operated by most major MNCs clearly exposes the deficiencies inherent in his proposal. He acknowledges that there is a preponderance of voluntary corporate codes adopted by major MNCs with respect to human rights,[57] and states that the regulatory initiative which he proposes at the institutional level will be voluntary in the sense that government will not enforce it, but maintains that the code will not be altogether without teeth. In his view, the teeth will be provided by the institutional compliance mechanisms put in place by the concerned MNC as well as the activities of stakeholders and civil society groups.[58] However, it is argued that there is nothing in the proposed code that makes it any different from other corporate codes. There is nothing new in his suggestion that the formulation of his proposed code will be guided by the content of regulatory initiatives at the international and national levels. The analysis of corporate codes by the European Commission reveals that most corporate codes are already guided by the contents of major international regulatory initiatives and human rights instruments.[59]

Secondly, civil society groups and other stakeholders already play prominent roles with regard to the implementation of corporate codes.[60] Deva's tenuous case for stakeholders to have a role in the implementation of his suggested code

[56] ibid 207.

[57] *Business and Human Rights Resourse Centre*, 'Company policy statements on human rights' <https://www.business-humanrights.org/en/company-policy-statements-on-human-rights> accessed 16 February 2016.

[58] Deva (n 8) 207.

[59] *European Commission*, 'An Analysis of Policy References Made by Large EU Companies to Internationally Recognised CSR Guidelines and Principles' (March 2013) 7 <https://ec.europa.eu/docsroom/documents/10372/attachments/1/translations/en/renditions/native> accessed 16 February 2016.

[60] Organisation for Economic Cooperation and Development, *Making Code of Conduct Work: Management Control System and Corporate Responsibility* (OECD Publishing 2001/03) 3 <http://dx.doi.org/10.1787/525708844763> accessed 16 February 2016.

contradicts his earlier view that the code does not necessarily have to conform to any national or international guidelines or incorporate the opinions of stakeholders.[61]

Thirdly, apart from the fact that the MNCs themselves are to formulate the proposed code, implementation of and compliance with the code will also depend on the adopting corporations. Given the focus of MNCs on profit, it is doubtful that they will put in place any compliance mechanism with the necessary teeth, especially if compliance with such a code will in any way affect their economic interest. Deva himself had earlier argued that "in the absence of a clear, positive relation of codes of conduct to business profits, several corporations may be hesitant to adopt and/or implement corporate codes; corporations will not regulate themselves into competitive disadvantage".[62]

How Deva's proposed code could have been useful to his case study, Bhopal, remains to be seen. The MNC that was supposed to design and adopt the code in this case, UCC, was not particularly interested in the successful running of the company because the Bhopal plant was considered unprofitable.[63] The failures that led to the Bhopal disaster did not result from the absence of a code but from economic considerations that led UCC to adopt inferior safety and maintenance standards. Therefore, it is argued that Deva's proposed code would have been unable to change the Bhopal situation.

B. A review of Deva's proposed regulation at the national level

Deva argues that regulation at the national level is an "indispensable medium to control and redress corporate human rights abuses",[64] but he does not think that the current regulations at such levels are enough to control and redress abuses. To achieve regulatory efficiency, he suggests that a regulatory regime at the national level should aim to influence corporate conduct both from the *outside* and the *inside*. He therefore suggests the revision of existing national laws that touch on human rights—such as labour law, investment law, environmental law, consumer protection law, *et cetera*—or the enactment of new laws to incorporate principles governing corporate human rights responsibilities.[65]

Deva's emphasis on an approach that involves changing corporate conduct from the inside is instructive. He criticises the external influence model's focus on the outcome of corporate decisions, which he argues merely specifies an outcome to be achieved on a given issue, and then responds with either a sanction or a

[61] Deva (n 8) 207.
[62] Deva (n 8) 78.
[63] ibid 28–29.
[64] ibid 28.
[65] ibid 209.

reward.[66] According to Deva, most laws that try to regulate corporate conduct in the area of human rights fall into this category, and he argues that this approach is insufficient to regulate human rights-related issues concerning MNCs.[67]

It is also instructive that corporate law is Deva's focus "regarding bringing about changes from the inside (*i.e.* in the process of corporate decision making)".[68] He argues that changes in corporate law are required because the premise on which the fundamental principles of the corporate law of all advanced economies are based has changed drastically, which makes it difficult to inject human rights responsibility into corporate decision-making.[69] This problem, according to him, arises from the uni-focal nature of the present corporate law (or practice), "conceiving corporations solely or primarily as profit maximizing entities, which puts pressure on corporate managers to pursue the goal of maximizing profit with total disregard for the interests of stakeholders other than shareholders".[70]

Deva argues that there are many approaches that could be adopted to bring about the proposed change; in fact, he cites some countries that have already amended their corporate law to broaden its scope to encompass the interests of other stakeholders beyond shareholders (such as the United Kingdom (UK), South Africa, and India).[71] The implication and extent of the obligation of corporate directors to other stakeholders under these regimes are discussed below.

Deva's reference to the way in which MNCs misuse the twin principles of separate legal personality and limited liability to evade their liability for human rights violations[72] is instructive. He acknowledges the importance of these principles and argues that they should not be allowed to become a standard refuge for corporate irresponsibility.[73] However, the problem is that his proposed approach to avoiding the misuse of the principles does not in reality reflect change from the inside. This is because he adopts the three *remedial responses*: (1) allowing case-by-case *ad hoc* exceptions to the twin principle; (2) the enterprise principle; and (3) the network liability approaches canvassed by Peter Muchlinski.[74]

It is important to point out that the responses canvassed by Muchlinski above are aimed at 'lifting the corporate veil' as a means of justifying group liability in circumstances where the subsidiary of a MNC has insufficient assets to meet the

66 ibid 208.
67 ibid.
68 Deva (n 8) 211.
69 ibid.
70 ibid.
71 ibid 196–197.
72 ibid 212.
73 ibid 213.
74 Muchlinski (n 1) 321–326.

claims against it,[75] as happened in the Bhopal case. These remedial solutions to corporate violations of human rights are at variance with Deva's proposed change of corporate conduct from the inside, which he acknowledges involves a review of the current corporate legal regime to broaden its scope to include the interests of other stakeholders.[76] The mismatch between, on the one hand, Deva's identified problem of the misuse of the twin principles of separate legal personality and limited liability, and on the other, his suggested solution to the problem deserves further treatment here because of its implication for the regulation of corporate violations of human rights.

There is an important point to note here about the modern corporation and its two guiding principles: separate legal personality and its close cousin, limited liability. Both features were developed to support the 19th century focus of nation states on the use of the corporation as a structure for promoting economic activities.[77] The legal personality principle, on the one hand, created the capacity for parties to act as a single entity in law—to sue and be sued, to hold and transfer title to real or personal property and to act with legal effect under a common seal.[78] The result is that the enterprise is able to continue undisturbed in law by a change of shareholders or a change in their fortune because, by virtue of the principle, corporate assets are shielded from both the shareholders and their creditors.[79] The limited liability doctrine, on the other hand, encouraged contribution from diverse investors by limiting the risk to their personal wealth, as limited liability provides a statutory assurance that "nobody risks more than he chips in".[80]

However, in this modern era of corporate group and multi-national operations by major MNCs, the twin concepts of separate personality and limited liability has brought about a new challenge. This is because of the extension of these concepts to corporate groups (a concept by which a MNC is permitted to hold the majority or all the shares in a subsidiary or subsidiaries), a phenomenon that developed after these principles were established.[81] This has generated much concern in

[75] ibid 313.

[76] Deva (n 8) 211–212.

[77] James W Hurst, *The Legitimacy of the Business Corporation in the Law of the United States 1780–1970* (The University Press of Virginia 1970) 22.

[78] ibid 19.

[79] John Armour and his colleagues termed the legal personality doctrine 'entity shielding' device "to emphasize that it involves shielding the assets of the entity—the corporation—from the creditors of the entity's owners". See John Armour et al, 'What is Corporate Law?' in Reinier Kraakman *et al*, *Anatomy of Corporate Law: A Comparative and Functional Approach* (3rd edn, OUP 2009) 5–8.

[80] Frank H Easterbrook and DR Fischel, *The Economic Structure of Corporate Law* (Harvard University Press 1996) 40.

[81] Phillip I Blumberg, *The Multinational Challenge to Corporate Law: The Search for a New Corporate Personality* (OUP 1993) 139.

the commercial world, as courts are often called upon to answer the question of whether those corporations which are members of a group are to be treated as a single economic entity (enterprise theory). In this regard, the courts have often rejected the idea that corporations in a corporate group be treated as a single economic entity.[82] The celebrated case of *Adams v Cape Industries Plc*[83] epitomises the judicial view that the 'veil of incorporation' cannot be disregarded in order to hold a parent company liable for the act or omission of the subsidiary.[84]

It is in light of judicial pronouncements, such as the dictum by Slade LJ in the *Adams* case above, that Muchlinski's suggestion is relevant as a strategy for attaching legal liability to a parent company for the acts or omissions of its subsidiary. In fact, Muchlinski criticised the UK Company Law Review Steering Group[85] for failing "to confront the question whether the *Adams* case went too far in blocking veil lifting in appropriate cases, such as where involuntary creditors needed to seek out the resources of the group as a whole for adequate compensation".[86]

Thus, it is for the purposes of securing adequate compensation for corporate tort victims that Muchlinski chose to base his analysis of corporate group liability on two competing objectives. One of those objectives is relevant to the present discussion: "the need to ensure that the resulting allocation of risk in the group does not end in a failure to compensate third parties for losses caused by the activities of group members".[87] In this regard, where there is an allegation of a human rights violation, the enterprise theory could be adopted to treat the whole group as one enterprise as this would satisfy the objective of proper compensation of third parties affected by human rights violations. It is in furtherance of this view, and to support his own suggestion of how to avoid the misuse of the separate legal personality principle, that Deva argues forcefully for the adoption of the theory of 'limited eclipse personality'. This theory is proposed to temporarily eclipse the separate legal personality of the subsidiaries of the corporate group in the case of alleged human rights violations, so that the victims are free to sue the immediate or ultimate parent corporation of that group as a matter of principle.[88]

In view of the above, even though this article agrees with Deva on the indispensability of national laws for controlling and redressing corporate human rights violations, it disagrees with his proposed remedial approach for the following

[82] See *Adams v Cape Industries Plc* [1990] BCLC 479, 513.
[83] ibid.
[84] ibid 513.
[85] The Company Law Steering Group, *Modern Company Law for a Competitive Economy* (London: Department for Trade and Industry, 2001).
[86] Muchlinski (n 1) 326.
[87] ibid 321.
[88] Deva (n 8) 213.

reasons. First, Deva himself acknowledges that command and control rules, which influence the corporate conduct from outside, by merely specifying an outcome to be achieved and responding with sanctions, are not sufficient.[89] This informs his proposed change from the inside, which does not merely "influence corporate decisions but also decision-making processes by changing the internal structure of corporations".[90] However, as the above discussion has shown, Deva's remedial-based solution does not address the unifocal nature of present corporate law and practice, and thereby fails to consider his proposed change from the inside that will provide a basis to influence corporate decisions and promote human rights integrity.

Secondly, Deva's reliance on the remedial approach considerably diminishes the opportunity for him adequately to address his proposed change from the inside, which is supposed to inject human rights responsibility into corporate decision-making.[91] In particular, his proposal is not consistent with the so-called amendment of the corporate legal regime in the UK, India, and South Africa, which—according to him—imposes duties on corporate managers to take the interests of non-shareholders into account. Under the new corporate law regimes in the countries he identifies above, the internal structure of the corporation is not in reality affected by the introduction of those provisions that, according to Deva, broadens the scope of corporate law to accommodate the interests of stakeholders.[92] This is because the changes he identifies do not provide for non-shareholders per se or to be part of the decision-making process, and no mechanism is provided through which non-shareholders can hold those who make the decisions to account.

For example, s 172 of the Companies Act 2006 (hereinafter, "UK CA 2006"), which Deva relied on to identify the UK as one of those jurisdictions which have made changes in their corporate law to accommodate non-shareholder stakeholder interests, is merely an embodiment of the concept of 'enlightened shareholder value' (ESV).[93] The ESV theory itself did not completely depart from the vision that directors are to manage the corporation in such a way as to ensure that the wealth of shareholders is maximised. The theory is achieved in the UK CA 2006 "through the high-level 'statement of directors' duties' set out in the Act[94] to clarify the duties and responsibilities of directors".[95] The theory maintains that the interest of shareholders is the principal obligation of directors and requires that directors

[89] Deva (n 8) 208.
[90] ibid.
[91] ibid 211.
[92] ibid 196–197.
[93] Jean Jacques du Plessis *et al*, *Principles of Contemporary Corporate Governance* (3rd edn, CUP 2015) 60.
[94] UK CA 2006 (n 19), s 172.
[95] du Plessis (n 93) 60.

pursue shareholders' interests but that, in doing so, they are to have regard to the interest of other stakeholders.[96]

In view of the above, it is argued that s 172 of the UK CA 2006 did not alter the traditional UK corporate governance structure that is designed to protect shareholders. In fact, some commentators take the view that s 172 is unlikely to strengthen the position of other stakeholders and thus suggest that fundamental changes to the current UK corporate legal framework would be necessary for them to have effective influence.[97] Thus, the system is inadequate for protecting the interests of non-shareholders because, absent the change from the inside which Deva proposes, non-shareholder constituencies will have to depend on those laws that influence the corporation from outside, which Deva himself criticises.

C. A review of Deva's proposed regulation at the international level

Another aspect of Deva's 'integrated approach' is an international framework that can formulate corporate human rights responsibility and ensure its implementation by MNCs as a way to overcome the limitations of national regulatory initiatives.[98] It is noteworthy that Deva issues a caveat on the numerous obstacles against such effort. According to him, "an agreement on developing international norms and international enforcement mechanisms is not proving to be an easy one in view of the numerous looming challenges".[99]

Deva premises his proposed regulation at the international level on international agreement about corporate human rights responsibilities, which he expects to be given a more precise meaning at the national level.[100] This raises some fundamental questions because international agreements are important sources of international law,[101] and national law is the medium through which States implement their obligations under international law.[102] This makes a review of the interaction between domestic law and international law—especially international human rights law—relevant. Understanding how issues of human rights and

[96] UK CA 2006 (n 19), s 172(1)(b), (c) and (d).

[97] John Birds et al (eds), *Boyle & Birds' Company Law* (9th edn, Jordan Publishing Limited 2014) 371–375.

[98] Deva (n 8) 214.

[99] ibid.

[100] ibid 215.

[101] Thomas Buergenthal and Sean Murphy, *Public International Law in a Nutshell* (4th edn, West Academic Publishing 2007) 24, citing the Statute of the International Court of Justice (ICJ Statute) (18 April 1946) Article 38(1)(b).

[102] Patricia Birnie, Alan Boyle and Catherine Redgwell, *International Law and the Environment* (3rd edn, OUP 2009) 270.

business corporations are addressed at the international level will provide the appropriate basis to determine how Deva's proposed international regulation could meaningfully address corporate human rights violations in developing markets.

International human rights instruments are devoted to recognising the individual's rights and the corresponding obligations of States. The International Covenant on Civil and Political Rights (ICCPR)[103] clearly states the obligation on each State party to the Covenant to respect and to ensure to all individuals within its territory and subject to its jurisdiction the rights recognised in the ICCPR.[104] However, the emergence of powerful non-state actors—especially MNCs—on the international stage has engendered new concerns about corporate accountability for human rights. This has brought about a renewed interest in the question whether there are real, direct legal obligations for non-state actors, especially MNCs, contained in international human rights instruments.[105] As a result, there has been an emerging broad interpretation of the rights enshrined in the International Bill of Rights (IBR)[106] in the context of corporate human rights concerns which were not there at the time existing rights were first formulated.[107]

Thus, in this context, while states are clearly the primary addressees of human rights obligations, corporations are also bound by those rules of international law that are applicable to all persons (natural and non-natural including business corporations) in society.[108] An example is the preamble to the Universal Declaration of Human Rights (UDHR) which states that "every individual and every organ of society should promote respect for basic human rights". Apart from the UDHR, both the ICCPR and the International Covenant on Economic, Social and Cultural Rights (ICESCR)[109] also recognise private obligations in their preambles, which is stated in the following terms: "the individual, having duties to other individuals

[103] International Covenant on Civil and Political Rights (ICCPR) (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171, 173.

[104] ibid Article 2(1).

[105] August Reinisch, 'The Changing International Legal Framework for Dealing with Non-State Actors' in Philip Alston (ed), Non-State Actors and Human Rights (OUP 2005) 70–71.

[106] "International Bill of Rights" is the expression used for the Universal Declaration of Human Rights, the ICCPR and the International Covenant on Economic, Social and Cultural Rights.

[107] Michael Anderson, 'Human Rights Approaches to Environmental Protection: An Overview' in Alan Boyle and Michael Anderson (eds), *Human Rights Approaches to Environmental Protection* (Clarendon Press 1996).

[108] Reinisch (n 105).

[109] International Covenant on Economic, Social and Cultural Rights (adopted 16 December 1966, entered into force 3 January 1976) 993 UNTS 3, 5.

and to the community of which he belongs, is under a responsibility to strive for the promotion and observance of the rights recognized in the present Covenant".

Even though corporations are bound by those rules of international law that are applicable to all persons in the IBR, "such rules are mostly restricted to fundamental norms such as those enjoining genocide, torture, slavery and forced labour, crimes against humanity,…".[110] In fact, "outside the European Union Framework, it is only in exceptional circumstances that corporations are expressly and directly regulated under international human rights law".[111] Related to this is the fact that there are only few mechanisms under international law[112] to enforce human rights norms against private actors outside the criminal sphere.[113] Thus, the responsibility for implementation and enforcement of international human rights norms against private actors such as MNCs lies primarily at the national level.[114] Accordingly, human rights standards can only be applied indirectly to corporations. In the case of MNCs operating in developing markets, any enforcement in relation to a breach of their obligation to respect human rights is either through the State in which they are incorporated (their home State)[115] or through the State in which they are operating (the host State).[116]

However, because of the obstacles to home States' regulation of foreign subsidiaries of locally incorporated companies, responsibility for domestic implementation and enforcement of international norms lies with the host States through their national laws and judicial institutions.[117] It is on this basis that some States, such as Nigeria, have been held responsible for their failure to prevent or remedy human rights abuses committed by private actors such as MNCs.[118] In the landmark case of *SERAC and CESR v Nigeria*,[119] the African Commission held that the obligation to protect human rights is a positive duty which requires States to

[110] Paul Redmond, 'Transnational Enterprise and Human Rights: Options for Standard Setting and Compliance' (2003) 37 Int'l L. 68, 71.

[111] ibid 72.

[112] See, for example, the ILO Tripartite Declaration of Principles concerning Multinational Enterprises and Social Policy (4th edn, Geneva: International Labour Office 2006), for non-binding interpretative procedures with respect to the Guidelines and Declarations that they have adopted concerning international business operations.

[113] Redmond (n 110) 72.

[114] ibid.

[115] *The Barcelona Traction Co* (Belgium v Spain) [1970] ICJ Rep 3.

[116] Redmond (n 110) 72.

[117] ibid.

[118] *Social and Economic Rights Action Centre (SERAC) and Centre for Economic and Social Rights (CESR) v Nigeria Communication* No 156/96 (Unreported, African Commission on Human and Peoples' Rights, 27 October 2001); *SERAP v Federal Republic of Nigeria*, Judgement ECW/CCJ/JUD/18/12 (14 December 2012).

[119] *SERAC v Nigeria* (ibid), para 59.

"take measures to protect beneficiaries of the protected rights against all political, economic and social interferences".[120]

The implication of the positive responsibility on States for human rights is that the implementation and enforcement of the private obligations recognised in both the preamble to the UDHR, the ICCPR and the ICESCR depends for the most part on a State's capability to protect individuals within its territory by adequately regulating every individual (including MNCs) and organs of society operating within its jurisdiction. For a State to fulfil its obligation to protect human rights requires that it not only provide adequate access to a remedy for human rights victims, but also take appropriate steps to prevent such violations through legislative or administrative means.[121]

There are several major problems identified as challenges to fulfilling the obligation to respect, protect, promote, and fulfil human rights for developing markets. First, the acknowledged economic power and influence of modern MNCs,[122] which affords them a power greater than some States to affect the realisation of rights.[123] Deva himself acknowledges how the economic power and influence of UCC played a major part in the manner in which the company operated in India, and how it contributed to the Bhopal disaster.[124] Secondly, the inefficient judicial infrastructure associated with developing markets is a major impediment to the implementation and enforcement of international norms, and renders any proposed regulation at the international level that focuses on providing remedies for human rights victims unsuitable, especially for developing markets.

In view of the fact that Deva expects his proposed international regulation to be given a more precise meaning at the national level, what institution will hold MNCs to account where they fail to comply with their human rights responsibilities under his proposed international framework? If municipal institutions will be required to enforce human rights responsibility under his proposed international

[120] Ibid, para 46.

[121] Dinah L Shelton, Advanced Introduction to International Human Rights Law (Edward Elgar Publishing Limited 2014) 214.

[122] See Sara Anderson and John Cavanagh, *Top 200: The Rising of Corporate Global Power* (New Press 2000), which concludes at 3 that "General Motors is now bigger than Denmark; DaimlerChrysler is bigger than Poland; Royal Dutch/Shell is bigger than Venezuela; IBM is bigger than Singapore; Sony is bigger than Pakistan", and that "top 200 corporations combined sales are bigger than the combined economies of all countries minus the biggest 10" at <https://corpwatch.org/downloads/top200.pdf> accessed 19 December 2017. See also John Gerard Ruggie, 'American Exceptionalism, Exemptionalism and Global Governance' (2004) John F Kennedy Sch of Govt, Working Paper No RWP04-006 1, 14, at <http:www.ssrn.com/abstract=517642> accessed 4 January 2016.

[123] John Gerard Ruggie, 'Business and Human Rights: The Evolving International Agenda' (2007) 101 AJIL 819, 824.

[124] Deva (n 8) 26–27.

framework, then the basis for Deva's international regulation is indeed faulty, as it will be unable to effectively support his integrated approach.

IV. A Critique Of 'Integrated Approach' And The Case For Interest Group Participation In Decision-Making

Deva's integrated approach—which requires a regulatory regime at three levels, each of which should complement the other in enforcing human rights responsibilities of MNCs[125]—makes sense if one considers the structure of MNCs and their global operation, but there are some major problems with Deva's suggested approach to integration. It is instructive that he strongly contends that corporations should have social responsibility. He argues that, as members of society, they are expected "to play their part by taking appropriate measures within their respective areas of operation: for example, ensuring that effluents from their factories are not discharged into rivers to preserve the environment".[126]

Deva is right in this respect because playing their part will include acting as responsible members of society and respecting the rights of others, such as local communities and the environment. However, to ensure that corporations behave responsibly will require some form of regulation as well as effective procedure for securing compliance. Thus, in view of the weak legal environment in developing markets, the focus here is on broadening the scope of corporate law to accommodate those that are most affected by corporate externalities as a strategy for promoting responsible corporate decision-making, and ensuring compliance with human rights standards. Deva himself agrees on the need to broaden the scope of corporate law to encompass the interests of other stakeholders beyond shareholders.[127] But he did not offer any suggestion as to how corporate law reform could form the basis for his suggested integration. It is argued that Deva's approach to integration is incomplete because he did not identify any platform at the national level for integrating international human rights standards.

In order to appreciate the uncertainty of Deva's approach to his suggested integration, it is important to offer a brief analysis of how he expects that the integrated application of the three levels of regulation might lead to a more responsible conduct by MNCs in developing markets. Deva adopts a 'top down' approach because, according to him, there is an emerging consensus on corporate human rights responsibilities at the international level which has gained recognition in various international instruments and major international

[125] Deva (n 8) 203.
[126] ibid 123.
[127] ibid 211–212.

regulatory initiatives,[128] such as the Organisation for Economic Cooperation and Development (OECD) Guidelines for Multinational Enterprises (MNEs)[129] and the UN Guiding Principles (UNHRC Guiding Principles).[130] This may be the case, but it is how this responsibility could be translated in the national context that is the issue.

Deva argues that regulatory initiatives at the institutional and national levels should conform to global human rights standards, but also suggests that some modifications to international human rights standards should be permitted at the national level to meet local circumstances.[131] However, he appears to contradict this position by suggesting that MNCs would be expected to apply home country or international standards in host countries. It is difficult to see how the permission to modify international human rights standards to *meet local circumstances*[132] will not negatively affect the application of home country or international standards in host countries, especially if it will be economically rewarding to apply a different (low) standard in a host country.

Another problem with Deva's approach to integration is that even though international norms and various regulatory initiatives may emphasise the responsibilities of businesses in some respects, his so-called 'top down' approach did not specify any process for the adoption of human rights responsibilities by businesses within the national or institutional context. He argues that a 'bottom up' approach will also be at work under his integration theory.[133] By this he means that the standards adopted by companies at an institutional level and those formulated by governments at a national level will form the basis for an agreement on standards at international level.[134]

It is clear inconsistency for the same author who argues that international regulatory initiatives would form the basis for regulatory standards at institutional and national levels to suggest that standards adopted and formulated at the same institutional and national levels will form the basis for agreements at the

[128] Deva (n 8) 202.

[129] Organisation for Economic Cooperation and Development, 'OECD Guidelines for Multinational Enterprises: Recommendations for Responsible Business Conduct in a Global Context' (25 May 2011) <https://www.oecd.org/corporate/mne/48004323.pdf> (hereinafter, "OECD Guidelines") accessed 4 January 2016.

[130] UNHRC, 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework: Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises' (21 March 2011) A/HRC/17/31.

[131] Deva (n 8) 202.

[132] Emphasis added.

[133] Deva (n 8) 202.

[134] ibid.

international level. Even though Deva argues that this is "[t]he crucial aspect of this integration process as a continuous upward-downward cycle of dialogue and evolution between regulatory initiatives at three levels",[135] it is doubtful that such 'top down' and 'bottom up' oscillations within his integration theory could provide an appropriate basis for a consistent regulatory regime for addressing corporate human rights abuses.

A. Interest group participation in decision-making as a strategy for integrating international norms at the national level and the basis for participation

It is agreed that international human rights standards, and other international norms and international regulatory initiatives should influence the contents of a national framework that aims to regulate corporate human rights violations in developing markets. However, it is argued that the focus of such regulation should be on how to prevent violations, rather than on how to provide effective remedies for victims. This makes it important to identify those international norms that espouse prevention as a basis upon which to build a national legal framework on corporate responsibility for human rights in developing markets.

The idea of a regulated corporate responsibility for human rights or what is commonly referred to as Corporate Social Responsibility (CSR) is not new. Campbell argues that "(CSR) thinking on corporate legal liability… can be taken to be established matters of fact in contemporary society".[136] This is not far from the prevailing position with regard to CSR and its regulation. Even the UNHRC Guiding Principles and the appended commentary that seek to implement the 'Protect, Respect and Remedy' Framework requires States to enact laws that "require, promote or guide companies to respect their human rights responsibilities".[137]

Thus, those developing markets where human rights violations are most prevalent should take steps to address the violations through legislation, including the amendment of their corporate laws. By broadening their scope so as to accommodate the interests of other stakeholders, States can fix the problem of linking corporate irresponsibility to the structure of corporate law.[138] This

[135] ibid.

[136] Tom Campbell, 'The Normative Grounding of Corporate Social Responsibility: A Human Rights Approach' in Doreen McBarnet, Aurora Voiculescu and Tom Campbell (eds), *The New Corporate Accountability: Corporate Social Responsibility and the Law* (CUP 2009) 529.

[137] UNHRC, 'New Guiding Principles on Business and Human Rights endorsed by the UN Human Rights Council' (16 June 2011) HR/PUBC/11/04 Principle 3.

[138] Robert C Hinkley, 'Developing Corporate Conscience' in Stuart Rees and Shelly Wright (eds), *Human Rights and Corporate Responsibility—A Dialogue* (Annandale: Pluto Press 2000) 287–290.

is because "[i]t is within the firm that the internal economic calculations and decisions are made—which do not include the external social costs or benefits that these decisions may impose outside the firm".[139]

In this regard, it is important to note that the kind of harm addressed in this article is not gratuitously inflicted for its own sake, as a means of deliberate repression or abuse, but emerges as a by-product of the scramble for economic development. It is for this reason that it is now widely accepted within the international legal order that development must be sustainable if it is not to prove as counterproductive as the Bhopal case and other such human rights disasters in developing markets. In the context of preventing environmental harm, international environmental law has developed approaches to prevention as elaborated in the extensive body of international environmental treaties[140] and related instruments[141] that explicitly and implicitly set out the principle of prevention with a focus on public and interest group participation.

The Rio Declaration on Environment adopted at the UN Conference on Environment and Development (UNCED) in 1992,[142] especially Principle 10, is regarded as the broadest and perhaps the most appropriate normative foundation for developing public and interest group participation at the national level as a strategy for preventing environmental harm.[143] It provides that "[e]nvironmental issues are best handled with the participation of all concerned citizens, at the

[139] Eric W Orts, *Business Persons: A Legal Theory of the Firm* (OUP 2015) 22.

[140] Ramsar Convention on Wetlands of International Importance 1971 (adopted 2 February 1991, entered into force 21 December 1975) 996 UNTS 245, 11 ILM 963; Convention Concerning the Protection of the World Cultural and Natural Heritage (adopted 16 November 1972, entered into force 17 December 1975) 1037 UNTS 151; Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal (adopted 22 March 1989, entered into force 5 May 1992) 28 ILM 657; United Nations Framework Convention on Climate Change (adopted 4 June 1992, entered into force 21 March 1994) (1992) 31 ILM 849; Convention on Biodiversity (adopted 6 June 1992, entered into force 29 December 1993) (1992) 31 ILM 818; United Nations Convention to Combat Desertification in those Countries Experiencing Serious Drought and/or Desertification, Particularly in Africa (adopted 14 October 1994, entered into force 26 December 1996) (1994) 33 ILM 1328.

[141] Declaration of the United Nations Conference on the Human Environment (June 1972) UN Doc A/CONF/48/14/REV.1; World Charter for Nature (adopted 29 October 1982) (1983) 22 ILM 455; Rio Declaration on Environment and Development (adopted 12 August 1992) (1992), adopted by the UN Conference on Environment and Development (UNCED) UN Doc A/CONF.151/26 (Vol I).

[142] UNGA 'Report of the United Nations Conference on Environment and Development' (12 August 1992) UN Doc A/CONF.151/26 (Vol I).

[143] Dinah L Shelton, 'Introduction' in Dinah L Shelton (ed), *Human Rights and the Environment Volume 1* (Edward Elgar Publishing Limited 2011) x.

relevant level. At the national level, each individual shall have… the opportunity to participate in decision-making processes".[144]

However, public participatory rights in decision-making process did not begin with their adoption in the international environmental law arena. International human rights law has long recognised civic participation in public affairs as a procedural right for ensuring the participation of those that may be affected by the decisions made by public authorities, in the decision-making process. The ICCPR guarantees the right of every citizen "[t]o take part in the conduct of public affairs, directly or through freely chosen representatives".[145]

It may be argued that international human rights law did not make detailed provision for prevention as a strategy for achieving human rights standards, and thus lacks the normative basis upon which to build public participation in the context of private sector operations. However, international environmental law, and especially the Rio Declaration, provides a normative basis for public and interest group participation especially in matters of economic development, which also involves MNCs. Although the Rio Declaration is 'soft law' and therefore has no binding authority in its own right, it has been suggested that environmental protection may be cast as a means to the end of fulfilling human right standards.[146] In this regard, existing human rights such as the right to life[147] and the right to a satisfactory environment[148] have been mobilised and reinterpreted as guaranteeing, to an extent, the public right to information and participation in decision-making processes.[149] This is because these are the very rights that are at issue in most developing markets.

In addition to the emphasis on public and interest group participation under international norms, emerging international regulatory initiatives—especially the OECD Guidelines and the UNHRC Guiding Principles—also emphasise the participation of potentially affected groups in decision-making concerning activities that are likely to have a serious human rights impact.[150] This is instructive

[144] Rio Declaration (n 142), Principle 10.
[145] ICCPR (n 103) Article 25.
[146] Anderson (n 107) 3.
[147] *Oneryildiz v Turkey* (2004) ECHR 657, paras 62 and 84–88.
[148] *SERAC v Nigeria* (n 118), para 41.
[149] Uzuazo Etemire, *Law and Practice on Public Participation in Environmental Matters: The Nigerian Example in Transnational Comparative Perspective* (Routledge 2016) 6.
[150] OECD Guidelines (n 123) para 14; UNHRC Guiding Principles (n 124) Principle 17.

because it comes from initiatives endorsed by the private sector.[151] The adoption of interest group participation in decision-making and the recognition of the authority of national governments to enact laws necessary to promote respect for human rights suggests that the regulatory target will be bound, or at least guided, by many of the principles expressed in such initiatives.

In view of the above, it is argued that what is currently required is an appropriate procedure for implementing the participatory principle as enunciated in the various international instruments, rather than the formulation of new human rights standards or specific international human rights norms for corporations as Deva suggests.[152] Thus, although Deva proposes regulations at three levels, it is argued that integration at the national and international levels could be usefully exploited in a manner that is different from his suggested approach in order to promote a more responsible corporate behaviour in developing markets. This could be achieved by exploiting State obligations under international instruments to develop a national legal framework that focuses on the participation of affected groups in decision-making concerning development activities that are likely to generate human rights harm. This will provide the opportunity for States to hold MNCs accountable to human rights standards.

B. Corporate law as the platform for integration and the case for local communities' representation on corporate boards

The inadequacy of existing legal and regulatory regimes that seek to address corporate human rights violations in developing markets is implicit in the growing cases of human rights violations in countries like Nigeria. Audrey Gaughran, Global Issues Director at Amnesty International, has described the human cost of such violations in the Niger Delta as horrific.[153] Some commentators have attributed this situation of impunity to the weak legal environment, especially the weak, corrupt and inefficient legal enforcement regimes, and point to the preference of human rights victims for foreign litigation, especially in the US under the Alien

[151] Both the OECD Guidelines and the UNGPs emerged as alternatives to influence UN's attempts at 'codification' to move away from highly regulatory position of MNC control as envisaged by both the UN Economic and Social Committee in 1974 (in the case of the OECD Guidelines), and the Economic and Social Council 1972 (in the case of the UNGPs). This provided the basis for the adoption of non-binding principles that could earn corporate as well as civil society support. See Muchlinski (n 1) 658–659; Deva (n 8) 106.

[152] Deva (n 8) 202.

[153] Amnesty International, 'Nigeria: Hundreds of oil spills continue to blight Niger Delta' (*Amnesty International*, 19 March 2015) <https://www.amnesty.org/en/latest/news/2015/03/hundreds-of-oil-spills-continue-to-blight-niger-delta/> accessed 13 December 2017.

Tort Claims Act (ATCA)[154] and more recently in the UK,[155] as a consequence of the weak legal environment in emerging markets.[156]

It is in view of the above that the participation of potentially affected groups in the affairs of the corporations operating in developing markets is proposed as a strategy for promoting responsible corporate decision-making and minimising the harm-creating potential of MNCs. It is argued that it makes sense to have this participation at the corporate board level because decisions of the board are necessary for undertaking the activities that create human rights abuses addressed in this article. Obviously, this is a question of corporate governance, since corporate governance is "the system by which companies are directed and controlled".[157] This makes it imperative to critically analyse the basis for the current corporate governance regimes' focus on shareholder value in order truly to understand how the underlying principles could usefully support the proposal for interest group participation in corporate affairs.

As matter of law, corporations must have a corporate board.[158] Apart from the German system with its two-tier board structure that permits employees representatives to sit on corporate boards,[159] corporate law in the UK (and in countries, including Nigeria and India,[160] that practice the UK's 'shareholder

[154] *Esther Kiobel v Royal Dutch Petroleum Co* (2013) 133 S Ct 1659; *Doe v Unocal* (2002) US App LEXIX 19263, *35–*36 (9th Cir 2002); *Wiwa v Royal Dutch Petroleum Co and Shell Transport and Trading Co* (2000) 226 F 3d 88.

[155] *His Royal Highness Emere Godwin Ebebe Okpabi and Others v Royal Dutch Shell Plc and Shell Petroleum Development Company of Nigeria Limited* [2017] EWHC 89 (TCC); *Dominic Liswaniso Lungowe and Others v Vedanta Resources Plc and Konkola Mines Plc* [2017] EWCA Civ 1528, [2018] WLR 3575; *Bodo Community v Shell Petroleum Development Company* (Nigeria) Ltd [2014] EWHC 1973 (TCC).

[156] Richard Meeran, 'Access to Remedy: The United Kingdom Experience of MNC Tort Litigation for Human Rights Violations' in Surya Deva and David Bilchitz (eds), *Human Rights Obligations of Business: Beyond the Corporate Responsibility to Respect* (CUP 2015) 382–383.

[157] See the definition of "corporate governance" in the *Report of the Committee on the Financial Aspect of Corporate Governance* (hereinafter, "Cadbury Report 1992") (Committee on the Financial Aspect of Corporate Governance, 1992); *The King Report on Corporate Governance* (hereinafter, "King 1994") (Institute of Directors in South Africa, 1994).

[158] See the UK CA 2006 (n 19), which prescribes 1 director in s 154(1), and a minimum of 2 directors for public companies in s 154(2). In Nigeria, The CAMA (n 19) prescribes a minimum of 2 directors in s 246(1).

[159] Johannes Schregle, 'Co-determination in the Federal Republic of Germany: A Comparative View' (1978) 117 Int'l Lab Rev 81, 83–99.

[160] Rafael La Porta *et al*, 'Law and Finance' (1998) 106(61) Journal of Pol Economy 1113, 1136.

supremacy' model[161]) focuses exclusively on the company and its shareholders.[162] In fact, under this model, corporate governance is considered effective if it provides mechanisms for regulating the exercise of directors' powers in order to restrain them from abusing their powers, just to ensure that they focus on their primary duty "to promote the success of the company for the benefit of its members as a whole".[163]

To ensure directors' fealty, only the shareholder group is granted powers under corporate law, at least in the 'shareholder supremacy' model jurisdictions, to control the board of directors or review its activities. The UK Companies Act 2006 (UK CA 2006) grants voting rights to shareholders,[164] including the right to vote on the appointment[165] and removal[166] of directors. In addition, it provides certain safeguards and control mechanisms for shareholders to rein in directors to ensure adequate protection for their class.[167] Under s 994(1) of the UK CA 2006, a member has powers to petition for relief against unfairly prejudicial conduct of the company's affairs. To reinforce shareholder protection, Part 30 (ss 994–999) of the Act gives the courts a very wide-ranging jurisdiction to remedy conduct

[161] Different corporate governance theorists classify corporate in various ways. However, the taxonomy that is relevant to our discussion here is that which categorises corporate governance according to the interests that the corporation serves, because of the distinction it makes between a corporate governance system that emphasises shareholder primacy (represented by the UK and the US), which is classified as the shareholder model, and the other end of the taxonomy, namely, a system that accounts for a wider group of constituents (represented by Germany), and classified as the stakeholder model. See the classification by Stephen Bainbridge (n 21) 8–9.

[162] Christine A Mallin, *Corporate Governance* (5th edn, OUP 2016) 178–184.

[163] UK CA 2006 (n 19), s 172.

[164] ibid ss 281–287. See also, in Nigeria's CAMA, ss 224–232.

[165] UK CA 2006 (n 19), s 160; CAMA (n 19), ss 247–249.

[166] UK CA 2006 (n 19), s 168; CAMA (n 19), s 262.

[167] Protection for creditors also exists under the CA 2006 and IA 1986. See Insolvency Act 1986, ss 213–214 with regard to director's responsibility to the company's creditors, in order to ensure that a company is not involved in fraudulent or wrongful trading. See also s 172(3) requiring directors, in certain circumstances, to consider or act in the interest of creditors of the company; and *West Mercia Safetywear Ltd v Dodd* [1988] BCLC 250. However, it is important to note that the sections cited above are only applicable when a company has entered formal insolvency proceedings and therefore differ from provisions which give control rights to shareholders. They are remedial rather than directional.

of a company's affairs "that is unfairly prejudicial to the interest of its members generally or of some part of its members".[168]

Notwithstanding all the rights and powers granted to members by statute, many commentators still doubt that shareholders could meaningfully exercise control over corporate boards in view of their dispersion, which has resulted in the effective separation of ownership from control in the firm, at least under the UK model.[169] However, MNCs and their subsidiaries present a different case. The power of the parent in the MNC to control the subsidiary or subsidiaries, is not in any doubt. As noted earlier, this is because of the way MNCs are usually structured.[170]

The common explanation for the vesting of voting and sundry rights in shareholders is that shareholders are the only corporate constituency with a residual claim, an unfixed, *ex post* claim on corporate assets and earnings.[171] This is the basis for the economic efficiency view, the one offered by Easterbrook and Fischel.[172] According to this view, because of shareholders' position as the last in the corporate revenue distribution chain, they "have the strongest economic incentive to care about the residual claim, which means that they have the greatest incentive to elect directors committed to maximizing firm profitability".[173] Many commentators, including this author, disagree with this view. In fact, some argue that it represents the "superficial analogy of the seventeenth century between contributors to a joint stock and members of a guild",[174] which no longer reflects the current reality.[175]

The focus of corporate boards on the shareholder constituency is primarily aimed at protecting it from the economic damage that may result from opportunistic behaviour by corporate managers.[176] This is perfectly in order but, in modern times, production processes do not pose only economic dangers. As a result of the increasing use of technologies and hazardous materials by modern corporations,

[168] See Part 30 (ss 994–999) CA 2006; members are sometimes permitted by the court to bring "derivative claims" in the name of the company. A "derivative claim" by a member of a company may be brought under CA 2006, Part 11 (ss 260–264). Relief on the grounds of unfairly prejudicial and oppressive conduct, derivative action and the powers of the court are also contained in CAMA (n 19), ss 310–333.

[169] Berle and Means, (n 23) 4–10; Bainbridge, (n 21) 4.

[170] See Mevorach (n 24).

[171] Bainbridge (n 21) 50.

[172] Easterbrook and Fischel (n 80) 66–72.

[173] Bainbridge (n 21) 50

[174] Abram Chayes, 'The Modern Corporation and the Rule of Law' in Edward S Mason (ed), *The Corporation in Modern Society* (Harvard University Press 1966) 41.

[175] Blumberg (n 81) ix.

[176] Charlotte Villiers, *Corporate Reporting and Company Law* (CUP 2006) 2, 18–19; French, Mayson and Ryan (n 20) 429.

especially those involved in mining and manufacturing, they also pose direct and serious threats to the health of communities surrounding their plants,[177] with far-reaching consequences.

Some of the most often-cited examples of human rights and environmental disasters that have occurred as a result of the neglect of the interests of other stakeholders include the blast at BP's Texas City Oil refinery which cost the lives of fifteen workers and injured 170 others; the Bhopal gas leak from the storage tank owned and operated by Union Carbide India Limited (UCIL), which claimed over 15,000 lives of the company's workers and local residents according to the 2003 official government report of the Bhopal Gas Tragedy Relief and Rehabilitation Department, State of Madhya Pradesh.[178] Others include the Bangladeshi factory building (Rana Plaza) collapse,[179] and the environmentally devastating oil spills in the Nigerian Niger Delta.[180]

As a result of the growing dangers that the modern corporation poses to its neighbours and other constituents, some commentators have argued that shareholders should no longer have a claim to priority consideration in corporate governance systems.[181] Current scholarship identifies other stakeholder groups affected by corporate activity to include (in no particular order) employees, suppliers, customers, creditors, local communities, governments and the environment.[182] However, the following analysis suggests that local communities that host the operating facilities of MNCs are the most vulnerable group with regard to the effects of corporate activity on human rights, and thus the group with the greatest incentive to promote more human rights-friendly corporate decision-making.

The emphasis on a special right for local communities in this article is based on two related factors that underpin the relationship between corporations and stakeholders. First, different relationships with the corporation have different characteristics. They may be voluntary or involuntary, and may be direct or indirect.[183] Relationships with shareholders and sundry stakeholders who do

[177] Ralph Nader, Mark Green and Joel Seligman, *Taming the Giant Corporation* (W W Norton & Company Inc 1976) 130.

[178] Amnesty International, *Clouds of Justice: Bhopal Disaster 20 Years On* (London: Amnesty International 2004) 12.

[179] 'Bangladesh Factory Collapse Toll Passes 1,000' *BBC* (London, 10 May 2013) <www.bbc.co.uk/news/world-asia-22476774> accessed 30 March 2017.

[180] See Human Rights Watch, 'The Price of Oil: Corporate Responsibility and Human Rights Violation in Nigeria's Oil Producing Communities' (*Human Right Watch*, January 1999) <http://www.hrw.org/reports/1999/nigeria/nigeria0199.pdf> accessed 26 April 2017.

[181] Marvin A Chirelstein, 'Corporate Law Reform' in James W McKie (ed), *Social Responsibility and the Business Predicament* (Washington DC: The Brookings Institution 1974) 67.

[182] Mallin (n 162) 74–77.

[183] Thomas Donaldson, Corporations & Morality (Englewood Cliff: Prentice-Hall Inc 1982) 32.

business with the corporation such as customers, creditors, suppliers and labour are mostly voluntary. The obligations that flow from such a relationship are direct and protected by statute or under contract.[184] Shareholders are protected under corporate law, insolvency law protects creditors, and customers and suppliers are protected under contract too.[185] Employment law, contracts and pension agreements protect labour. Relationships with local communities are the direct opposite. They are mostly involuntary and obligations that flow therefrom are indirect because they are not necessarily specified under statute or under contract.[186]

Secondly, relationships with the corporation could be determinable or permanent. Voluntary relationships with corporations are always determinable. For example, a shareholder who is dissatisfied with the activities of a corporate board may take the exit option by selling his shares, and a worker is free at any time to resign his appointment. However, the relationship between the corporation and local communities is mostly of a permanent and inescapable nature.[187]

In view of the above, and given the foundational role of government in providing the legal infrastructure for corporations, it might be useful to exploit the potential of corporate law to protect those most affected by corporate externalities. Achieving this will require a review of the corporate legal architecture to incorporate such interests, especially those of local communities in the corporate governance structure, thereby making them part of corporate decision-making process.

The suggestion for corporate law reform to incorporate interests other than shareholders in the corporate governance structure is not entirely new. The recognition and general attention paid to other stakeholders in corporate law statutes and regulations both nationally[188] and internationally[189] points to the growing consensus on the new understanding of the corporation as affecting a wide range of people beyond just shareholders. In fact, some continental European jurisdictions have long recognised the important stake that other stakeholders,

[184] Donaldson (n 183) 32.

[185] It is acknowledged that other forms of regulations such as consumer protection laws also protect customers.

[186] Donaldson (n 183) 32.

[187] Andrew Crane, Dirk Matten and Laura J Spence (eds), *Corporate Social Responsibility: Readings and Cases in a Global Context* (2nd edn, Routledge 2014) 292.

[188] See, for example, the UK CA 2006 (n 19), s 172; the Indian CA (n 19), s 135; and the South African Companies Act 2008, s 7.

[189] See, for example, the OECD Guidelines (n 123) para 44; and the UNHRC Guiding Principles (n 124) Principle 17.

especially employees, have in the corporations and provides for their participation in corporate governance.

The German "Co-determination" regime with its two-tier board structure of a supervisory board and a management board[190] is the common reference regime for external stakeholder participation in management at board level in Continental Europe. The German system permits at least one stakeholder group—labour—to sit on German corporate boards alongside the directors.[191] Thus, the external stakeholder model envisaged in this article is mirrored in the German concept of other stakeholders' participation on corporate boards as a strategy for ensuring that issues concerning them is given due attention at corporate board level.

However, since MNCs operate in developing markets through their subsidiaries, the new framework will be primarily concerned with those subsidiaries, and perhaps the MNC as a shareholder. This is based on the fact that incorporation in our reference jurisdictions is mandatory for corporations that seek to undertake any business activity in those countries. For example, s 54 of CAMA permits a MNC to incorporate a subsidiary in Nigeria to carry on its business functions. Thus, the proposed framework does not regulate the MNC *per se*, except with regard to its involvement in the subsidiary. The underpinning principle of the proposed framework is the balance of interests: business corporations' right to operate without unlawful interference and to make profit, and their host communities' right to a healthy environment and respect for and protection of their human rights.

It is expected that the proposed regime will be criticised just like the German co-determination regime, especially with regard to its effect on other stakeholders and on a company's business policy. For example, in their suggestion for board restructuring to accommodate other stakeholder interests, Nader *et al* did not propose interest groups participation on corporate boards primarily because of what they considered a major defect of the German Co-determination regime. According to them, "[l]abor representatives apparently are indifferent to most problems of corporate management that do not affect labour. They seem as deferential to operating executive management as present American directors are".[192]

Although it is not possible to undertake a more detailed critique of the German co-determination because of the limited space allowed for this research, it suffices to state that there are studies that rebut the claims by Nader *et al* and other critics. In a 1976 survey of European experience with industrial democracy

[190] Co-determination Act, Law of May 1976, (1976) BGBI 1 1153.
[191] Mallin (n 162) 20.
[192] Nader, Green and Seligman (n 177) 124.

and worker representation at board level prepared for the Bullock Committee,[193] Batstone observed that German labour directors do not take advantage of their board position to redefine corporate objectives or challenge management strategy but generally support programmes such as investment that will strengthen the position of the company.[194]

However, despite the criticisms, the importance of co-determination as a strategy for promoting co-operation within the enterprise cannot be overemphasised. Bringing labour within the management structure and making their interests an integral part of the company satisfies the underlying idea of co-determination, which is to provide a rational means of handling and settling disputes at the enterprise level.[195] This is especially important to the discussion here, given that this article has identified the participation of local communities representatives at board level as a strategy for guaranteeing that issues concerning the human rights of MNC host communities are given due attention at board level. There is a clear symmetry between this strategy and the principles that underpin the co-determination regime. It is on this basis that I propose to explore the option of a corporate legal framework that draws on the principles of co-determination by developing a corporate board participatory regime involving local community representatives in corporate decision-making, especially for developing markets.

V. Conclusion

One major point of agreement that this article shares with Deva is on the impact that the growing power and influence of MNCs and their operations in developing markets is having on the human rights of local communities, and the urgent need to fashion a workable regulatory regime to adequately address the problem. National laws are currently incapable of regulating a difficult regulatory target like a MNC because of their domestic focus amongst other deficiencies. However, this article's point of disagreement is that corporate law in individual jurisdictions has the inherent potential to adequately regulate MNC conduct and thus prevent corporate violations of human rights.

Even though it is agreed that there are many factors that promote corporate irresponsibility in developing markets, it is nevertheless instructive that Deva

[193] The Bullock Committee was set up 3 December 1975 by the then Secretary of States for Trade, Peter Shore, 'to consider ways of extending industrial democracy in the private sector of industry' in the UK. See the preface to Eric Batstone and Paul Davies, *Industrial Democracy: European Experience* (The Stationery Office 1976).

[194] Eric Batstone, 'Industrial Democracy and Worker Representation at Board Level: A Review of the European Experience' in Eric Batstone and Paul Davies, *Industrial Democracy: European Experience* (The Stationery Office 1976) 25.

[195] ibid.

acknowledges that the change in corporate structure is fundamental to the present situation of corporate impunity for human rights violations. He also agrees that municipal laws and institutions are important even though he insists that they are incapable of effectively regulating MNCs. However, the fact that Deva did not specifically address how re-modelling of corporate law could address the fundamental problem posed by the current corporate structure that underscores his focus on a remedial solution.

The problem with the focus on a remedial solution is that the weaknesses associated with judicial institutions in developing markets makes a regime that focuses on remedy (and not prevention) an unreliable option. It also makes a human rights approach that is based on Deva's strategy of integrated regulation an unreliable option too. Therefore, it is argued that developing markets require a regime that accords with available corporate support institutions (which include securities law, securities regulations, securities enforcement and, the one most relevant to external stakeholders—public enforcement, such the courts) within the jurisdiction. In this regard, a legislative option with a focus on prevention, which emphasises public participation in corporate decision-making based on international norms, could be formulated to provide a more reliable basis to regulate MNC conduct. This will provide the opportunity for potentially affected groups, especially the local communities in developing markets to hold MNCs accountable to human rights standards.

However, because of the limited space allowed for this research, it is not possible to fully develop the suggested model in greater detail. Therefore, further research is recommended to develop the details of the procedure which would be regulated by the corporate legal regime.

Towards Greater Legal Protection for Medical-Humanitarian NGOs in Situations of Armed Conflict

Elizabeth Rose Donnelly*

La médecine, comme la souffrance, ne connaît pas de frontière.[1]

I. Introduction

For decades, international non-governmental organisations (NGOs) have provided emergency medical care to victims of war. Guided by the principle that those who do not (or no longer) participate in armed hostilities have the right to "medical assistance wherever and whoever they are",[2] NGOs such as Médecins Sans Frontières (MSF) and Médecins du Monde[3] have entered conflict zones on medical-humanitarian missions. They are afforded the protection of international

* M.A. (Cantab), LL.B. (Hons.), LL.M. My thanks to Dr Stephen Humphreys and Dr Louise Arimatsu at the London School of Economics and Political Science for their guidance and support.

1 "Medicine, like suffering, knows no borders." Adapted from Maurice Torrelli, 'La protection du médecin dans les conflits armés' in Christophe Swinarski (ed), *Studies and essays on international humanitarian law and Red Cross principles* (Martinus Nijhoff 1984) 585.

2 Yves Beigbeder, *The Role and Status of International Humanitarian Volunteers and Organizations* (Martinus Nijhoff 1991) 347.

3 See Médecins Sans Frontières UK <www.msf.org.uk/> accessed 10 February 2018; Médecins du Monde <www.medecinsdumonde.org/en> accessed 10 February 2018.

humanitarian law (IHL)[4] provided, *inter alia*, that they practise non-discrimination and refrain from interfering in States' internal affairs.[5] The norms regulating humanitarian relief and guaranteeing personal protection acknowledge their non-combatant status.[6] This additional legal protection is not a 'homage' to the medical profession,[7] but recognises that proximity to combat zones when treating victims necessitates a "more specific protection than that afforded the civilian population in general".[8]

However, humanitarians' security has been significantly weakened in recent years. Warring parties increasingly threaten staff, harry aid convoys, and target medical personnel and facilities during armed attacks.[9] In 2015 alone, aerial and ground shelling in Syria killed or wounded 81 medical staff in 63 hospitals and clinics—all supported by MSF.[10]

Two issues are germane to this. The first issue is that the very nature of armed conflict has evolved. Whereas inter-State armed conflicts once formed the principal focus of IHL, internal conflicts between government forces and rebel armed groups (or 'insurgents') now abound. This rise in asymmetric warfare engenders

[4] The principal sources are the four Geneva Conventions (GC) (GC/I: Convention for the Amelioration of the Wounded and Sick in Armed Forces and Field (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31; GC/II: Convention for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31; GC/III: Geneva Convention Relative to the Treatment of Prisoners of War (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31; GC/IV: Geneva Convention Relative to the Protection of Civilian Persons in Time of War (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 31) and their two 1977 Additional Protocols (AP/I: Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3; AP/II: Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of Non-International Armed Conflicts (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3). Also see Adam Roberts and Richard Guelff, *Documents on the Laws of War* (3rd edn, OUP 2000) 2, 5–7, 10.

[5] Reginald Moreels, 'Humanitarian diplomacy: The essence of humanitarian assistance' in Frits Kalshoven (ed) *Assisting the Victims of Armed Conflict and Other Disasters* (Martinus Nijhoff 1989) 43.

[6] Kate Mackintosh, 'Beyond the Red Cross: The protection of independent humanitarian organizations and their staff in international humanitarian law' (2007) 89(865) IRRC 113, 118.

[7] Jean-Pierre Schoenholzer, 'Le médecin dans les Conventions de Genève de 1949' (1954) 35(410) IRRC 94, cited in Torrelli (n 1) 591.

[8] Mackintosh (n 6) 123.

[9] Rebecca Barber, 'Facilitating humanitarian assistance in international humanitarian and human rights law' (2009) 91(874) IRRC 371, 373.

[10] Médecins Sans Frontières, 'Syria 2015: Documenting war- wounded and war- dead in MSF- supported medical facilities in Syria' (*Médecins Sans Frontières*, 8 February 2016) <www.msf.org/sites/msf.org/files/syria_2015_war-dead_and_war-wounded_report_en.pdf> accessed 22 July 2018.

violation—or 'reinterpretation'—of IHL;[11] the consequent "toxic atmosphere of defiance of law and order"[12] arguably leaves States with a greater propensity to reject international humanitarian assistance.

The second issue is one of access. Under IHL, an offer of international humanitarian assistance must be accepted by a State before missions can enter its lands. Authorisation is often delayed or arbitrarily (and unlawfully) withheld[13]—particularly when a government wants to keep humanitarians out of insurgent-controlled territory.[14] The response by 'no-borders' NGOs, and particularly by MSF, is to refuse to accept State consent as a *sine qua non* of humanitarian access.[15] As private actors, they lack international status,[16] and so cannot violate the principles of State sovereignty and non-intervention.[17] However, unauthorised missions forfeit various IHL protections.[18] MSF may argue that it acts "in advance of a constantly changing International Law"[19] and its call for missions to have full and unconditional respect may well be justified, but the "grim reality"[20] is that States cede IHL protection only *when* and *if* it suits them.

This article examines the phenomenon of unauthorised, wartime missions by international, medical-humanitarian NGOs (MNGOs), and its effect on their international legal protection. The binary question this article seeks to answer is whether (a) 'no-borders' organisations should desist from such missions, and "find ways to achieve their aims within the existing legal… structure" of IHL;[21] or whether (b) as MSF believes, the pursuit of unconditional, international legal protection remains valid. Part II presents MNGOs' humanitarian 'credentials' and the effect of these on qualification for IHL protection. Following Part III's overview of the relevant IHL protection provisions, Part IV offers two case-studies which examine the contemporary challenge of securing humanitarian access and protection for MNGOs' unauthorised missions in non-international armed conflicts. After considering the limits of IHL in this context, Part V analyses alternative means

[11] Marco Sassòli, 'The implementation of international humanitarian law: Current and inherent challenges' (2007) 10 Yearbook of International Humanitarian Law 45, 58–59.

[12] Yoram Dinstein, *Non-international Armed Conflicts in International Law* (CUP 2014) 115.

[13] See Dapo Akande and Emanuela-Chiara Gillard, 'Arbitrary withholding of consent to humanitarian relief operations in armed conflict' (2016) 92 International Law Studies 483; Cedric Ryngaert, 'Humanitarian assistance and the conundrum of consent: A legal perspective' (2013) 5(2) Amsterdam Law Forum 5, 9–11.

[14] Beigbeder (n 2) 347–348.

[15] Moreels (n 5) 47.

[16] Beigbeder (n 2) 327.

[17] Ryngaert (n 13) 12–13.

[18] International Committee of the Red Cross (ICRC), 'Customary International Humanitarian Law Database, Rule 55. Access for Humanitarian Relief to Civilians in Need' <https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule55> accessed 22 July 2018.

[19] Beigbeder (n 2) 348.

[20] Michael Meyer, 'Humanitarian action: A delicate balancing act' (1987) 27(260) IRRC 485, 497.

[21] ibid 497.

of securing enhanced legal status and protection for such MNGOs. This article concludes that, like armed groups, 'no-borders' MNGOs are now key players in situations of armed conflict, but the limits of their protection under IHL have been reached. In consequence, it is time that the international community works to develop international law in such a way that MNGO personnel are afforded unconditional protection and access to victims of armed conflict.

II. Impartial Humanitarian Organisations and International Humanitarian Law

A. Medical-humanitarian ngos in international humanitarian law treaties

As there is no settled definition of 'international humanitarian assistance' in international law, this article draws on Kalshoven's reference to *emergency* assistance from *outside* a country: "activities and goods which, out of a feeling of solidarity and joint responsibility, are designed to provide direct support to the victims of an armed conflict".[22]

In the context of armed conflict, international medical assistance is the provision to non-combatants of hospital supplies and equipment, medicines, vaccines, and skills of professional medical volunteers (including doctors and nurses).[23] Although international (or 'external') humanitarian assistance should *supplement* that provided by national authorities, it sometimes becomes entirely substitutive when authorities cannot, or will not, assist vulnerable citizens.[24] Some MSF missions have built and funded health facilities in regions where none exist, and paid the salaries of local medical staff. In 2014, it supported 56 Syrian medical facilities in regions impenetrable to international staff.[25] By 2016, this had risen

[22] Kalshoven (n 5) 20.

[23] ICRC, 'Respecting and Protecting Health Care in Armed Conflicts and in Situations Not Covered by International Humanitarian Law' (ICRC, March 2012) <https://www.icrc.org/eng/assets/files/2012/health-care-law-factsheet-icrc-eng.pdf> accessed 22 July 2018.

[24] Mario Bettati, 'La contribution des organisations non-gouvernementales à la formation et à l'application des normes internationales' in Mario Bettati and Pierre-Marie Dupuy (eds) *Les O.N.G. et le Droit International* (Economica 1986) 21.

[25] Sophie Delaunay, 'Condemned to resist' (*Professionals in Humanitarian Assistance and Protection*, 10 February 2014) 5, <https://phap.org/system/files/article_pdf/Delaunay-CondemnedToResist.pdf> accessed 22 July 2018.

to around 70, principally run by local doctors,[26] in Syria's southern and north-western regions.[27]

Medical assistance should be provided according to the principles of humanity, impartiality and non-discrimination.[28] The principle of impartiality comprises "the recognition of equality of all people, the duty of equal treatment, and… appropriate relief without favour or prejudice".[29] In the *Nicaragua* case, the International Court of Justice (ICJ) went further, stipulating that external humanitarian actors should support both or all parties to a conflict.[30] However, Kalshoven disagrees with this finding and considers that it goes against Red Cross principles (to which MNGOs also adhere).[31] Barrat suggests that impartiality "does not necessarily mean mathematical equality",[32] especially in situations where humanitarians may not be authorised to access all parts of a State's territory (for example, during a non-international armed conflict).

B. Medical-humanitarian NGOs' qualification for international humanitarian law protection

The Geneva Conventions (GCs) and their Additional Protocols (APs) were drafted prior to MNGOs' rise in prominence, and do not expressly provide for their operation and protection. It is accepted, however, that treaty references to "impartial humanitarian organizations"[33] or "some humanitarian organization"[34] encompass MNGOs.[35] Insofar as the treaties suggest that organisations must be *independent* (citing the International Committee of the Red Cross (ICRC) as an

[26] Zena Tahhan, 'MSF: Attacks on aid groups part of Syrian regime plan' *Al Jazeera* (Doha, 10 October 2016) <http://www.aljazeera.com/news/2016/10/msf-attacks-aid-groups-part-syrian-regime-plan-161010062509695.html> accessed 22 July 2018.

[27] 'Syria: Mapping the conflict' *BBC* (London, 10 July 2015) <http://www.bbc.co.uk/news/world-middle-east-22798391> accessed 22 July 2018.

[28] UNGA Res 182, 'Strengthening of the Coordination of Humanitarian Emergency Assistance of the United Nations' (19 December 1991) UN Doc A/RES/46/182.

[29] Claudie Barrat, *Status of NGOs in International Humanitarian Law* (Brill Nijhoff 2014) 148.

[30] *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America)* (Merits) [1984] ICJ Rep 392 [115].

[31] Frits Kalshoven, 'Impartialité et neutralité dans le droit et la pratique humanitaires' (1989) 71(780) IRRC 541, 548.

[32] Barrat (n 29) 150.

[33] Articles 9, 9, 9, and 10 of the GC/I, GC/II, GC/III, and GC/IV (n 4) respectively; also see GC/IV (n 4), Articles 59(2) and 61(1).

[34] GC/IV (n 4), Article 15.

[35] Mackintosh (n 6) 115–116.

example),[36] it seems that those maintaining financial and political independence can claim specific Convention protections.[37]

However, in interpreting the humanitarian principle of *neutrality*, there has been a conscious departure from the ICRC's position.[38] Having long since come to typify "the values of humanitarian universalism",[39] the ICRC practises neutrality not only by refusing to 'take sides', but also by refusing to criticise warring parties.[40] At times, the latter has been interpreted as indifference.[41] For French doctors witnessing large-scale atrocities and killings in the 1967–1970 Biafra conflict, the ICRC's silence was anathema to humanitarianism: they left to found MSF.[42] The MSF Charter declares that "principles of impartiality and neutrality are not synonymous with silence", and that its personnel "may speak out publicly [if witness to] extreme acts of violence… unacceptable suffering… [or when] medical facilities come under threat".[43]

Médecins du Monde goes further still in its commitment to "bear witness"[44] to atrocities, not infrequently 'taking sides' in conflict zones and denouncing the acts of warring parties.[45] Therefore, it is a matter of contention whether 'no-borders' neutrality—undertaking not to "take sides or intervene according to the demands of governments or warring parties"[46] but maintaining "freedom of criticism"[47]—breaches the conditions of humanitarian missions.[48] Beigbeder suggests that it

[36] The ICRC was not created by international treaty, qualifies as an NGO under the law of neutral Switzerland, and receives no funding from governments: see Beigbeder (n 2) 64–68.

[37] Mackintosh (n 6) 116.

[38] Note that the ICRC can function as a Protecting Power (a neutral intermediary between belligerent parties in an international (inter-State) armed conflicts (IACs)) if no State is able or willing to do so. See Articles 9, 9, 9, and 10 of the GC/I, GC/II, GC/III, and GC/IV (n 4) respectively; AP/I (n 4) Article 5(4)). The issue lies beyond this article's scope; for further, see Christophe Swinarski, 'La notion d'un organisme neutre' in Swinarski (ed) (n 1) 826–834.

[39] David Chandler, 'The road to military humanitarianism: How the human rights NGOs shaped a new humanitarian age' (2001) 23(3) Human Rights Quarterly 678, 679.

[40] ibid 684.

[41] Jacques Meurant, 'Principes fondamentaux de la Croix-Rouge et humanitarisme moderne' in Christophe Swinarski (ed) *Studies and essays on international humanitarian law and Red Cross principles* (Martinus Nijhoff 1984) 899.

[42] Rony Brauman, 'Médecins Sans Frontières and the ICRC: matters of principle' (2012) 94(888) IRRC 1523, 1524–1525.

[43] Médecins Sans Frontières, 'Who we are – The MSF Charter' <https://www.msf.org/who-we-are> accessed 22 July 2018.

[44] Médecins du Monde, 'Our Fundamentals' <http://www.medecinsdumonde.org/en/our-values> accessed 10 February 2018.

[45] Beigbeder (n 2) 266–267.

[46] Médecins Sans Frontières (n 43).

[47] Chandler (n 39) 685.

[48] Meyer (n 20) 495.

forfeits IHL protection of missions and personnel (though, as will be discussed, relevant personnel should remain entitled to protection as civilians).[49] Conversely, Plattner argues that the *legal* notion of neutral humanitarian assistance is "not dependent on the nature of the body" providing it;[50] if organisations' actions are impartial and non-discriminatory, the act of denouncing does not automatically forfeit IHL protection.[51] Barrat even suggests that, because neutrality was not identified by the ICJ in the *Nicaragua* case as essential to humanitarian assistance, it is not intrinsic to organisations' legitimacy.[52]

In reality, warring parties do not tolerate external medical-humanitarian actors who criticise, even those manifesting Plattner's conception of neutral action. The risk is that government forces might penalise or even attack MNGOs on the pretext that their public denunciations constitute internal interference.[53] In 1984, for example, a French MSF mission had to re-enter Ethiopia clandestinely after being removed for denouncing authorities' forced displacement of rural civilians.[54]

III. The Law on Protection and Access in Armed Conflict

A. The law applicable to armed conflicts

Today's medical-humanitarian NGOs operate amid a complex web of treaty provisions and customary IHL. The laws of war place armed hostilities within a "bifurcated legal framework".[55] International (inter-State) armed conflicts (IACs) are regulated by the Geneva Conventions and their First Additional Protocol (AP/I). Non-international (internal) armed conflicts (NIACs) commonly involve non-State insurgent groups and are regulated by common Article 3 of the four

[49] Beigbeder (n 2) 346–347.

[50] Denise Plattner, 'ICRC neutrality and neutrality in humanitarian assistance' (2006) 36(311) IRRC 161, 178.

[51] ibid 178–179.

[52] Barrat (n 29) 154.

[53] René Jean Dupuy, 'L'assistance humanitaire comme droit de l'homme contre la souveraineté de l'état' in Frits Kalshoven (ed) *Assisting the Victims of Armed Conflict and Other Disasters* (Martinus Nijhoff 1989) 33.

[54] Beigbeder (n 2) 264–265.

[55] Kenneth Watkin, '21st-century conflict and international humanitarian law: Status quo or change?' in Schmitt and Pejic (eds) *International Law and Armed Conflict: Exploring the Faultlines* (Martinus Nijhoff 2007) 267.

Geneva Conventions and, in certain circumstances, the Second Additional Protocol (AP/II).[56]

Unlike the Protocols, which are not universally ratified, the Geneva Conventions apply to all States. Common Article 3 of the four Geneva Conventions constitutes customary international law,[57] and it is generally accepted that certain rules of IAC-related law have also entered custom,[58] thereby enabling their application in NIACs.[59]

Yet, it is not always clear which provisions apply in a theatre of hostilities. The nature of contemporary conflict is such that some jurists speak of inter-State conflict "as a disappearing if not extinct concept".[60] There have been a few IACs in recent years (notably the 1999 Kosovo campaign and 2003 invasion of Iraq),[61] but most conflicts are NIACs. Moreover, the International Criminal Tribunal for the former Yugoslavia (ICTY) has previously stated that an IAC and NIAC can exist within the same territory.[62] For example, at the time of writing, Syria's armed forces are fighting against a US-led international coalition and various insurgent armed groups.[63]

B. International humanitarian law protection available to medical-humanitarian NGOs

In terms of the cardinal IHL principle of "distinction",[64] former and non-participants in hostilities must be protected from attack (this also extends to civilian

[56] ibid, 267–271. See also ICTY, *Prosecutor v Tadic* (Appeals Chamber) Judgment, IT-94-I-A, 15 July 1999 [84]: the ICTY declared that a NIAC becomes an IAC if "another State intervenes in that conflict through its troops' or 'some of the participants in the [NIAC] act on behalf of that other State".

[57] A "minimum yardstick" reflecting "elementary considerations of humanity": see Nicaragua (n 30) [218].

[58] Yoram Dinstein, *The Conduct of Hostilities under the Law of International Armed Conflict* (3rd edn, CUP 2016) 16–17: As a general rule, customary international law is applicable to all States. Custom is constituted from State-practice and *"opinio juris"*, that is, the "[States'] belief that this practice is rendered obligatory by the existence of a rule of law requiring it." (See *North Sea Continental Shelf (Federal Republic of Germany/Denmark)* (Merits) [1969] ICJ Rep 3 [44]).

[59] Watkin (n 55) 273.

[60] ibid 269.

[61] ibid 269–270.

[62] ICTY, Tadic (n 56) para 84.

[63] Geneva Academy, 'Syria' (Rule of Law in Armed Conflict Project, 14 February 2018) <http://www.rulac.org/browse/countries/syria> accessed 22 July 2018: "Syria is currently engaged in a series of armed conflicts. First, the Syrian government is engaged in several non-international armed conflicts against a wide array of rebel groups. Secondly, there is arguably an international armed conflict between Syria and members of the US-led international coalition and Turkey."

[64] Dinstein (n 58) 12.

objects).[65] Although this is "intransgressible" under customary law,[66] the protection of MNGO missions and personnel in conflict zones can be unclear. Therefore, the international community has sought in recent years "to improve respect for the normative framework assumed to protect" them.[67] This has yielded introduction of the Red Crystal (a 'culturally neutral' protective emblem, identical in function to the Red Cross and Red Crescent symbols identifying medical personnel *attached to a warring party*) and an Optional Protocol to the Convention on Safety of UN and Associated Personnel.[68] MNGOs can only benefit if their missions are carried out under the aegis of the UN, or qualify for use of a protective emblem by their relinquishing political independence.[69] Those that eschew all such associations, such as MSF, must continue to rely on the far-from-perfect legal protection provided by the Geneva Conventions.[70]

(I) LEGAL PROTECTION DURING INTERNATIONAL ARMED CONFLICTS

During IACs, as a customary rule, volunteers, staff and facilities of impartial humanitarian organisations benefit from general, civilian immunity from attack.[71] MSF has previously suggested that its volunteers and staff qualify for additional protection as "civilian medical personnel",[72] and that "IHL protects the legal autonomy of the medical mission within the mandatory rules of medical ethics pertaining to that profession".[73] Of course, if the relevant belligerent State authorises the medical mission, MNGOs should have no need to claim this 'supplementary' protection: belligerent parties are obligated to "respect and protect" medical-humanitarian missions' units, transport and personnel.[74]

However, for MNGO personnel in the hands of a warring party, IHL is less straightforward: protection turns on the individual's nationality.[75] The Fourth Geneva Convention provides protection according to non-combatant status, but

65 ibid 72.

66 *Advisory Opinion on the Legality of the Threat or use of Nuclear Weapons* [1996] ICJ Rep 2.

67 Mackintosh (n 6) 113.

68 ibid 113–114.

69 ibid 113, 124–125.

70 Note that MNGO personnel are also entitled to protection under international human rights law, e.g. the right to life (Article 6) and freedom from torture (Article 7) of the ICCPR (International Covenant on Civil and Political Rights 1966 (adopted 19 December 1966, entered into force 23 March 1976) 999 UNTS 171).

71 Mackintosh (n 6) 118.

72 Françoise Bouchet-Saulnier, 'Coutume: espace de création et d'activisme pour le juge et pour les organisations non-gouvernementales' in Tavernier and Henkaerts (eds) *Droit international humanitaire coutumier: Enjeux et défis contemporains* (Bruylant 2008) 169, cited in Barrat (n 29) 191.

73 Françoise Bouchet-Saulnier, 'Consent to humanitarian access: An obligation triggered by territorial control, not States' rights' (2014) 96(893) IRRC 207, 212–213.

74 AP/I (n 4) Articles 9(2)(c), 12(2), and 71.

75 Mackintosh (n 6) 118–119.

largely covers "civilians in the hands of the adversary"[76]—nationals of an "enemy State", or a country without diplomatic ties with that State.[77] As Mackintosh notes, MNGOs looking to preserve their neutrality and impartiality might (like the ICRC) send only non-nationals of State-parties to the conflict, but those volunteers would have next to no protection under the Fourth Convention (excepting the diplomatic representation caveat).[78] The alternative—sending relief workers who *are* nationals of a warring State—guarantees Fourth Convention protection, but jeopardises the MNGO's perceived neutrality[79] (note, however, that *all* MNGO personnel qualify for the more limited protection under Article 75 of the first Additional Protocol, which provides that persons who, when under a belligerent Party's power, do not "benefit from more favourable treatment under the Conventions… shall be treated humanely in all circumstances").

Meanwhile, protection for MNGO personnel acting as *civilian medics* theoretically stems from belligerents' "basic obligations to respect the wounded and sick".[80] The wounded and sick must either be civilians or persons placed *hors de combat* who "refrain from acts of hostility".[81] The first Additional Protocol states that "[n]o one shall be harmed, prosecuted, convicted or punished for… humanitarian acts", nor "punished for carrying out medical activities compatible with medical ethics, regardless of the person benefiting therefrom".[82]

(II) LEGAL PROTECTION DURING NON-INTERNATIONAL ARMED CONFLICTS

All NIACs are covered by common Article 3 of the four Geneva Conventions.[83] Whether internal armed violence constitutes a NIAC thereunder is established "on a case-by-case basis".[84] Generally, hostilities must reach a certain level of 'intensity', with non-State armed groups constituting organised armed forces in possession of a command structure, and with the capacity to conduct military operations.[85] The threshold for applying the Second Protocol is significantly higher: a government's armed forces must be involved, and the non-State adversary must exercise "such

[76] Roberts and Guelff (n 4) 299.
[77] Article 4 GC/IV; Mackintosh (n 6) 119.
[78] Mackintosh (n 6) 119.
[79] ibid.
[80] Alexander Breitegger, 'The legal framework applicable to insecurity and violence affecting the delivery of health care in armed conflicts and other emergencies' (2013) 95(889) IRRC 83, 107–108. See also GC/I (n 3) Article 12 and GC/II (n 4) Article 12; GC/IV (n 4) Article 16; AP/1 (n 4) Article 16.
[81] AP/I (n 4) Article 8(a).
[82] AP/I (n 4) Articles 17(1) and 16(1).
[83] Dinstein (n 12) 20.
[84] ICTR, *Prosecutor v Rutaganda* (Trial Chamber) Judgment, ICTR-96-3-T, 6 December 1999 [95].
[85] ICTY, *Prosecutor v Tadic* (Trial Chamber) Judgment, IT-94-1-T, 7 May 1997 [561]–[568].

control over a part of [the State's] territory as to enable them to carry out sustained and concerted military operations and to implement this Protocol".[86]

The *level* of territorial control required for an armed group to implement the second additional Protocol is a matter of debate. The International Criminal Tribunal for Rwanda (ICTR) in *Akayesu* held that insurgents "must be able to dominate a *sufficient* part of the territory so as to" conduct military operations and implement the Protocol.[87] The deciding factor is quality, not quantity,[88] that is, "effective territorial control".[89]

MNGO personnel, as non-combatants, should fall within the protective scope of common Article 3 of the four Geneva Conventions. In NIACs occurring within a State-party to the second Additional Protocol, such protection is enhanced by Article 4 and includes, *inter alia*, prohibitions on "violence to life [and] health" and "outrages upon personal dignity". Volunteers should be also protected because they tend the wounded and sick;[90] they cannot be punished for medical activities.[91] In theory, this should apply even when authorisation to the mission has not been granted in accordance with Article 18 of the second Additional Protocol. Volunteers detained by a warring party are owed humane treatment, regardless of nationality.[92]

(III) LOSS OF PROTECTION FOR A MEDICAL-HUMANITARIAN NGO'S PERSONNEL

Civilian medical personnel lose their IHL protection by committing an act outside their humanitarian function or which could "harm the adverse party, by facilitating or impeding operations".[93] Crucially, MNGOs should guard against

[86] AP/II (n 4) Article 1(1).

[87] *Prosecutor v Akayesu* (Trial Chamber) Judgment, ICTR-96-4-T, 2 September 1998 [626].

[88] Sandesh Sivakumaran, *The Law of Non-international Armed Conflict* (OUP 2012) 186; Dinstein (n 12) 45.

[89] Emily Crawford, *The Treatment of Combatants and Insurgents under the Law of Armed Conflict* (OUP 2010) 163.

[90] AP/II (n 4) Article 7.

[91] AP/II (n 4) Article 10.

[92] AP/II (n 4) Article 5.

[93] ICRC, Commentary to Article 21 of the GC/I (n 4) <https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=859FF3DB19BCAC7BC12563CD-00420FE1> accessed 22 July 2018.

actions which might threaten belligerents' perception of their impartiality and neutrality.[94]

IV. Case Studies: Problems of Access and Protection for Medical-Humanitarian NGOs in Non-International Armed Conflicts

Despite the range of legal protections available in IACs, it is submitted that NIAC-related law provides relatively scant protection, thereby posing a great challenge to MNGOs operating in the context of internal conflict. This is exacerbated when a warring party arbitrarily withholds consent to external assistance, as the following two case-studies will show.

A. When the territorial armed group consents to external medical aid, but the state does not

Entering territory under the *de facto* control of a non-State armed group without State consent has led to medical humanitarians being accused of 'siding' with rebels, such that their impartial status is deemed lost. During the Biafra conflict (1967–1970), for example, the ICRC defied the Nigerian Government's demand for aid to be transported through land corridors. When an unauthorised ICRC relief plane was shot down over Nigerian airspace, the Government accused the organisation of acting as "an agent for the [Biafra] secessionists".[95] More recently, MSF has been engaged in humanitarian missions in war-torn Syria. The Government repeatedly has refused to authorise the MNGO's work, criminalising any medical activity outside of government control.[96] Nevertheless, in 2012, MSF clandestinely opened a newly-built medical centre in the rebel-held north, smuggling equipment into the territory from neighbouring countries.[97] The MNGO maintained that its missions were valid under *international* law: "Maybe we were illegal for the Syrian regime, but at least we were legitimate".[98]

As noted above, the second Additional Protocol does not apply to all NIACs. When considering 'no-borders' MNGOs' access to a non-consenting State's

[94] For further discussion, see Breitegger (n 80) 110–111.

[95] Sivakumaran (n 88) 334.

[96] Kareem Shaheen, 'MSF stops sharing Syria hospital locations after "deliberate" attacks' *The Guardian* (London, 18 February 2016) <https://www.theguardian.com/world/2016/feb/18/msf-will-not-share-syria-gps-locations-after-deliberate-attacks> accessed 22 July 2018.

[97] Ruth Sherlock, 'Syria: Médecins Sans Frontières' secret hospital' *The Daily Telegraph* (London, 21 August 2012) <http://www.telegraph.co.uk/news/worldnews/middleeast/syria/9490764/Syria-Medecins-Sans-Frontieres-secret-hospital.html> accessed 22 July 2018; Tahhan (n 26).

[98] Tahhan (n 26).

sovereign territory, it is important to recall the differing legal protection when the State is not a party to this Protocol.

(I) TERRITORY OF A STATE WHICH HAS NOT RATIFIED ADDITIONAL PROTOCOL II

In territory controlled by non-State adversaries, there is a "divergence of views" regarding consent to humanitarian missions.[99] Common Article 3(2) of the four Geneva Conventions states that an "impartial humanitarian body… may offer its services to the Parties to the conflict"; this 'right of initiative' clearly also applies to MNGOs.[100] However, there is no provision as to *which* Party's consent is needed.[101] On the one hand, various jurists consider that humanitarian operations can be authorised by a non-State Party, provided that the territory it effectively controls can be accessed without entering onto that controlled by the non-consenting government.[102] Such was MSF's interpretation of IHL applicable to the Syrian conflict in the context of medical personnel's direct cross-border entry into rebel-held territory. And, as Syria is not party to the second Additional Protocol,[103] only common Article 3 would apply to MSF missions in a NIAC context. Conversely, Akande and Gillard indicate that common Article 3(2) cannot be interpreted to mean that *other States* may undertake humanitarian missions without State-party consent, as this would entail "the significant infringement of territorial sovereignty".[104] NGOs, as private actors, cannot be bound by the cardinal international law principles of sovereignty and territorial integrity[105]—an explanation which Bouchet-Saulnier offers as a possible reason for the Syrian government's manipulation of "domestic legal provisions converting medical relief into a weapon of war".[106]

[99] Dapo Akande and Emanuela-Chiara Gillard, *Oxford Guidance on the Law Relating to Humanitarian Relief Operations in Situations of Armed Conflict, (The United Nations Office for the Coordination of Humanitarian Affairs and Oxford University*, October 2016) 16 <https://www.law.ox.ac.uk/content/oxford-guidance-law-relating-humanitarian-relief-operations-situations-armed-conflict> accessed 22 July 2018.

[100] Maurice Torrelli, 'From humanitarian assistance to "intervention on humanitarian grounds"?' (1992) 32(288) IRRC 228, 231.

[101] Akande and Gillard (n 99) 16.

[102] See, for example, Michael Bothe, 'Relief Actions: The position of the recipient state' in Frits Kalshoven (ed) *Assisting the Victims of Armed Conflict and Other Disasters* (Martinus Nijhoff 1989) 94; Torrelli (n 100) 233–234; Bouchet-Saulnier (n 73) 210–211.

[103] ICRC Database, Parties to Protocol II <https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/States.xsp?xp_viewStates=XPages_NORMStatesParties&xp_treatySelected=475> accessed 22 July 2018.

[104] Akande and Gillard (n 99) 17.

[105] Ryngaert (n 13) 12–13.

[106] Bouchet-Saulnier (n 73) 213.

(II) TERRITORY OF A STATE-PARTY TO ADDITIONAL PROTOCOL II

For medical-humanitarian missions undertaken in the territory of a party to the second Additional Protocol, the question of State consent is even murkier. Article 18(2) states:

> If the civilian population is suffering undue hardship owing to a lack of supplies essential for its survival… relief actions for the civilian population which are of an exclusively humanitarian nature and which are conducted without any adverse distinction shall be undertaken subject to the consent of the High Contracting Party *concerned* [emphasis in italics added].

A literal reading suggests that a State's consent is *always* necessary for aid provision throughout its territory: it is 'concerned' by any humanitarian mission occurring there.[107] According to Sivakumaran, this renders the consent obtained from rebel groups insufficient because of the "insufficient attention paid to the specificities" of NIACs during the Protocol drafting process.[108] Consent from such insurgents exercising *de facto* control over State territory, in his view, is clearly a practical necessity for humanitarian actors.[109] And Bothe *et al* suggest that, as a State cannot be 'concerned' by operations occurring in territory over which it has no effective control, humanitarians may enter rebel-held territory provided they do not cross that of the State (as argued above in relation to common Article 3 of the four Geneva Conventions).[110]

Gillard challenges this interpretation of Article 18(2) of the second Additional Protocol, arguing that the decision to remove the phrase "*the party or parties concerned*" from the final draft demonstrated States' intention to exclude rebel groups from having power to consent.[111] However, Gillard does not analyse the silence of Article 18(2) of the second Additional Protocol on the legal status of humanitarian actors. It is submitted here that the provision implicitly refers to external relief provided by *States*. The Commentary reflects this: consent is not solely "left to the discretion

[107] Akande and Gillard (n 99) 17.

[108] Sivakumaran (n 88) 332.

[109] ibid.

[110] Michael Bothe, Karl Josef Partsch and Waldermar Solf, N*ew Rules for Victims of Armed Conflicts – Commentary on the Two 1977 Protocols Additional to the Geneva Conventions of 1949* (Martinus Nijhoff 1982) 694.

[111] Emanuela-Chiara Gillard, 'The law regulating cross-border relief operations' (2013) 95(890) IRRC 351, 365–366.

of the parties".[112] "Relief actions must take place" in territory where civilians are "suffering undue hardship"[113] if an impartial humanitarian organisation "is able to remedy this situation".[114] On this interpretation, MNGO missions qualify for protection under the second Additional Protocol in armed groups' territory—even in the absence of State consent.

No clear solution exists for MNGOs operating in rebel-held territory without State authorisation; they must stake their safety on warring parties interpreting relevant IHL norms in good faith, respecting civilian and medical-humanitarian immunity. To complicate matters further, MNGOs like MSF insist on adherence to *medical* neutrality. If the needs of a wounded combatant are greatest, s/he will be treated first.[115] In light of the 'civilians-only' stipulation of Article 18(2) of the second Additional Protocol, a warring State-party to the Second Protocol might deem an MNGO's impartiality and neutrality defunct on learning of enemy combatants being treated.

B. When the territorial armed group does not consent to external medical aid

Without an armed group's consent to external humanitarian assistance, it is almost impossible for an external MNGO to carry out unauthorised missions in territory under that group's effective control. In such circumstances, the humanitarian toll can be staggering. Between 2012 and 2016, for example, Syrian government forces besieged and bombarded rebel-held Darayya. Without access to humanitarian assistance (to which neither side consented),[116] approximately

[112] International Committee of the Red Cross, Commentary to the Additional Protocols, para 4885 <https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=086657E594BB4CC2C12563CD0043ADD0> accessed 22 July 2018.

[113] AP/II (n 4), Article 18(2).

[114] Commentary to the Additional Protocols (n 112), para 4885.

[115] A senior MSF staff-member has criticised States for challenging medics' "right and duty to treat everyone, including combatants": see Kareem Shaheen, 'Hospitals are now targets of war, says Médecins Sans Frontières' *The Guardian* (London, 1 June 2016) <https://www.theguardian.com/world/2016/jun/01/hospitals-are-now-normal-targets-of-war-says-medecins-sans-frontieres-adviser> accessed 22 July 2018.

[116] ICRC, 'Syria: Aid convoy turns back after being refused entry to besieged Daraya: Joint statement by the ICRC and the UN' (ICRC, 12 May 2016) <https://www.icrc.org/en/document/aid-convoy-turns-back-after-being-refused-entry-besieged-daraya>; UN News Centre, 'Syria: UN Agencies Reach Families with Food in the Besieged Town of Darayya' (9 June 2016) <http://www.refworld.org/docid/575e59e940c.html> accessed 22 July 2018.

8,000 civilians were trapped without adequate food or medicine, with dozens dying through starvation or illness.[117]

(I) FEASIBILITY OF THE EXTENSION OF THE LAW OF BELLIGERENT OCCUPATION TO TERRITORIES CONTROLLED BY ARMED GROUPS

Clearly, for a population no longer under a State's *de facto* control, the legal protection-gap is significant within the context of a NIAC. As a somewhat radical solution, Gal has proposed extending norms of the international law of belligerent occupation to territory under insurgents' effective control.[118] Occupation law—a 'branch' of the laws of war[119]—currently regulates only IAC situations in which one State's armed forces take effective control of territory in an enemy State after "combat stabilizes along fixed lines … not coinciding with the original international frontiers".[120] Principally, this legal regime regulates "a trilateral relationship between the Occupying Power, the displaced sovereign and the civilian population of the occupied territory".[121] Its "cornerstone"[122] is constituted by the 1907 Hague Regulations[123] (now part of customary international law). Article 42 of the Regulations provides: "[T]erritory is considered occupied when it is actually placed under the authority of the hostile army. The occupation extends only to the territory where such authority has been established and can be exercised."

For civilians in occupied territory, specific provisions of the Fourth Geneva Convention and First Protocol guarantee "enhanced protection",[124] including a suite of obligations on the Occupying Power regarding humanitarian assistance.[125] Crucially, the law of occupation obviates the requirement of consent to external humanitarian assistance.[126] Article 59 of the fourth Geneva Convention imposes

[117] Hugh Naylor, 'In a Syrian town under a brutal siege, a young girl is left deaf and hopeless' *Washington Post* (Washington DC, 20 June 2016) <https://www.washingtonpost.com/world/middle_east/in-a-syrian-town-under-a-brutal-siege-a-young-girl-is-left-deaf-and-hopeless/2016/06/19/d65ccbc0-27e5-11e6-8329-6104954928d2_story.html?utm_term=.31e078d48413> accessed 22 July 2018.

[118] Tom Gal, 'Territorial control by armed groups and the regulation of access to humanitarian assistance' (2017) 50(1) Israel Law Review 25, 27.

[119] Yoram Dinstein, *The International Law of Belligerent Occupation* (CUP 2009) 3.

[120] ibid 1.

[121] ibid.

[122] ibid 4–6.

[123] Hague Regulations Respecting the Laws and Customs of War on Land, annexed to Hague Convention (IV) 1907 (adopted 18 October 1907, entered into force 26 January 1910).

[124] Dinstein (n 119) 6–7.

[125] ibid 194.

[126] GC/IV (n 4), Article 59.

an "unconditional"[127] obligation on the Occupying Power to grant access and protection to humanitarian relief for the civilian population if the territory is "inadequately supplied".[128] The occupant should be informed of planned missions, but consent (or lack thereof) is irrelevant. Gal suggests that these provisions could apply, *mutatis mutandis*, to territory under the effective control of non-State actors.[129] She argues that had occupation law extended to insurgents and applied to Darayya, the rebels' effective control would have obliged them to admit aid into the city. Syrian government authorities would also have been obliged to provide safe passage for aid convoys transiting through government-controlled territory.[130] In theory, since Article 59 permits "impartial humanitarian organisations" to provide assistance, MNGOs would be able to undertake missions in this context.

However, this thesis is arguably not without doctrinal obstacles. Occupation law applies only to territory controlled by an enemy *State* during an IAC.[131] Common Article 2(2) of the four Geneva Conventions requires that relevant Convention (and, now, first Additional Protocol) provisions apply to territory occupied during *international* conflict. In its *Wall* Opinion, the ICJ also confirmed the law's application to "territory occupied… by one of *the contracting parties*".[132] Consequently, it seems legally impossible to extend the occupation regime to NIACs.[133]

Yet, Gal argues that the "factual circumstances" of armed groups' territorial control should transcend the law's preoccupation with such groups' legal status or with State sovereignty.[134] The notion of effectiveness in international law can cause "a factual situation [to] strongly affect legal norms",[135] as the ICTY recognised: "[IHL] is a realistic body of law, grounded on the notion of effectiveness… [it] holds accountable not only those having formal positions of authority but also those who wield *de facto* power".[136] In short, since the effective control required of a State for occupation law to apply is analogous to that required of insurgents to

[127] ICRC, Commentary on GC/IV (n 4), Article 59 <https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/Comment.xsp?action=openDocument&documentId=15B5740DF2203BE4C12563CD0042C966> accessed 22 July 2018.

[128] GC/IV (n 4), Article 59.

[129] Gal (n 118) 27.

[130] Nevertheless, the authorities would have had the right to inspect aid consignments, to be 'reasonably satisfied' that relief provided to Darayya's civilian population would not 'be used for the benefit' of enemy belligerents (GC/IV (n 4), Articles 59(3)–(4)).

[131] Barber (n 9) 384–385.

[132] *Advisory Opinion Concerning Legal Consequences of the Construction of a Wall in Occupied Palestinian Territory* [2004] ICJ Reports 136 [95] (emphasis added).

[133] Torrelli (n 1) 599.

[134] Gal (n 118) 27.

[135] ibid 40.

[136] ICTY, *Prosecutor v Tadic* (Appeals Chamber) Judgment, IT-94-1-A, 15 July 1999 [96].

trigger the Second Protocol's application to (relevant) NIACs,[137] Gal suggests that extending occupation law to such internal conflicts would reflect "the spirit of this body of law", ensuring that "political aspirations and interests will not diminish the rights and needs of victims of war".[138]

Reference to the ICJ's *Wall Advisory Opinion* is absent from Gal's analysis. Perhaps her proposition is more in tune with academic support for extending occupation law to UN-controlled territory.[139] In this, Roberts observes that UN forces and non-State armed groups are similar: IHL treaties apply to them, even though they are not party thereto.[140] Common Article 3[141] of the four Geneva Conventions and Article 1(4) first Additional Protocol[142] already expressly cite armed groups as *addressees* of IHL. Ferraro notes not only that various UN peacekeeping and enforcement missions have exercised "functions and powers over a territory that could be compared to those assigned to an occupant under occupation law",[143] but also that the law of occupation is arguably "the only body of law capable of addressing the tension between the suspended sovereign and the new administering authority",[144] thereby facilitating continued civilian protection. Gal's argument about insurgent groups also reflects that of a panel of ICRC experts who agreed that UN forces' *effective control* of territory would be the key factor in applying occupation law.[145]

(II) Arguments against "extending" the law of belligerent occupation to rebel-held territory during non-international armed conflict

First, any *legal* solution such as Gal's would have little traction with States. It may well be that "[i]nternational reality… is less and less state-centred",[146] but there is still little or no incentive for States to extend occupation law to insurgent-controlled territory. Despite the enhanced protection guarantees for civilians,

[137] Gal (n 118) 42; Bothe (n 102) 94.
[138] Gal (n 118) 47.
[139] Adam Roberts, 'What is a military occupation?' (1985) 55(1) BYIL 249, 289.
[140] ibid.
[141] Sassòli (n 11) 63.
[142] This provision refers to national liberation movements 'fighting against colonial domination', 'alien occupation' and/or 'racist régimes'. The full raft of GC and AP/I (n 4) provisions apply to such conflicts. See Gal (n 118) 38–39, 43–44.
[143] Tristan Ferraro (ed), 'Expert meeting: Occupation and other forms of administration of foreign territory' (ICRC, March 2012) 33 <https://www.icrc.org/eng/assets/files/publications/icrc-002-4094.pdf> accessed 22 July 2018.
[144] ibid.
[145] ibid.
[146] Sassòli (n 11) 63.

States would likely baulk at insurgents' acquiring any status which might lend them administrative legitimacy, for the Occupying Power is permitted, in occupation law, to exercise its jurisdiction and conduct itself as "substitute for the legal sovereign".[147] It is simply not credible that Iraq or Syria, for example, would have accepted the extension of occupation law to parts of their territory controlled by so-called Islamic State.[148]

Secondly, an occupying power's territorial control can fluctuate, both in terms of effectiveness and geography. Dinstein notes that "loss of effective control as a result of defeat in the field may not last long, inasmuch as the pendulum of military ascendancy in war may swing again in the opposite direction".[149] One seasoned MSF aid-worker has compared the realities of the Syrian conflict with power struggles between insurgents and State forces in Mali:

> Timbuktu is emblematic of the need for impartial care as front lines shift… If one of our Syrian hospitals currently located in a rebel-controlled area would end up being located in government-held territory… our medical support would be… valuable for them…[150]

Whilst such a scenario has the potential to disrupt MNGOs' work, Gal's proposed NIAC scenario might engender even greater uncertainty.[151] An armed group's loss of effective territorial control would precipitate the loss of occupation protection guarantees to medical-humanitarian missions,[152] replacing them with the relative paucity of protection available in the law relating only to NIACs, and compromise the ability of MNGO personnel to establish medical facilities and enjoy free operational movement within the territory. The risk would be exacerbated by belligerent occupation, like armed conflict, being a question of

[147] Gal (n 118) 37.

[148] 'Islamic State and the crisis in Iraq and Syria in maps' *BBC* (London, 28 March 2018) <http://www.bbc.co.uk/news/world-middle-east-27838034> accessed 22 July 2018.

[149] Dinstein (n 119) 272; GC/IV (n 4), Article 70(1).

[150] Delaunay (n 25) 3.

[151] Gal notes that the International Criminal Court (ICC) referred to armed groups' control of Timbuktu as an 'occupation'; she interprets this as attaching 'further international obligations' to the insurgents. (See above, n 118, 46). Nevertheless, it is submitted here that the Court used "occupation" as a *non-legal* descriptor. See ICC, *Prosecutor v Ahmad Al-Faqi Al-Mahdi*, Decision on the Confirmation of Charges, ICC-01/12-01/15, Pre-Trial Chamber, 24 March 2016, [44]–[45], [55].

[152] An Occupying Power is obliged to permit 'medical personnel of all categories… to carry out their duties' and to respect and protect new hospitals established in the territory, in additional to medical convoys/transport. Civilian medics must be given 'every assistance'. See GC/IV (n 4) Articles 18, 20, 21, 56; AP/I (n 4) Article 15(3)

fact.[153] Even if an armed group respected the law of occupation, its analysis of events on the ground could diverge dramatically from that of the sovereign State or humanitarian organisation. This *ex post facto* determination of occupation—"whether this degree of control was established at the relevant times and in the relevant places"[154]—offers little consolation for medical-humanitarians whose IHL protection waxes and wanes with a NIAC's shifting contours.

V. Beyond International Humanitarian Law: Exploring Legal-Protection Options

Preventing or alleviating human suffering is a fundamental principle of IHL.[155] No-borders MNGOs, seeking to uphold that principle through medicine, venture into conflict zones despite decades of mistreatment of their personnel. Most of the international community has progressed from dismissing these organisations as "mercenaries in white coats",[156] but States remain ambivalent about granting them enhanced protections. Like armed groups, MNGOs are now key players in situations of armed conflict, but they remain "largely non-existent for international law".[157]

The requirement of State consent to external assistance in IHL remains the principal doctrinal (and practical) obstacle to protection for MNGOs in conflict zones. As Dinstein observes, "as long as consent is essential… authorities can usually find plausible excuses for delaying humanitarian assistance, and even for frustrating it altogether".[158] It is submitted, therefore, that the limits of IHL protection have been reached. MNGOs unwilling either to put up with States' arbitrary withholding of consent or to work under the aegis of UN or third-party States' military personnel, brave warring parties' capricious observance of IHL. Whilst jurists endeavour to develop new approaches in this area (as examined above), that corpus of law remains mired in States' competing interests.

This does not mean, however, that the need for greater legal protection has gone unnoticed by the international community. Various Resolutions of the

[153] Tristan Ferraro, 'Determining the beginning and end of an occupation under international humanitarian law' (2012) 94(885) IRRC 133, 134–135.

[154] ICTY, *Prosecutor v Naletilić & Martinović* (Trial Chamber) Judgment, IT-98-34-T, 31 March 2003 [218].

[155] Dupuy (n 53) 29.

[156] Torrelli (n 1) 600.

[157] Sassòli (n 11) 63.

[158] Yoram Dinstein, 'The right to humanitarian assistance' (2000) 53(4) Naval War College Review 77, 85.

Council of Europe and the UN Security Council[159] have addressed different aspects of international law in this regard. Time and again, however, there has been little or no meaningful progress, principally because, on a practical level, the provisions have been piecemeal and have not thoroughly dealt with rights of access and State consent. The final section of this article reflects on the merits and shortcomings of legal-protection options which have been proposed for MNGOs over the years.

A. Early draft instruments

At a Council of Europe conference in 1984, a draft Charter for the Protection of Medical Missions was proposed.[160] Several 'no-borders' MNGOs initiated the project, seeking recognition by States of the dangers faced by medical-humanitarian personnel in conflict zones.[161] The draft Charter reaffirmed several IHL norms (including the proscription of punishment for medical activities, respect for medical ethics, and free operational movement within a territory), and formulated various rights and obligations for civilians and medical humanitarians alike.[162] These included the right of civilians to be treated by competent medical professionals, the right of medical personnel to protection during missions, and the creation of an identifiable symbol or professional badge, to be ascribed by the ICRC.[163]

In the wake of this Charter proposal, the Council of Europe produced Resolution 904 (1988) "on the protection of humanitarian medical missions". A non-binding legal document, it advocated a rights-based approach to healthcare provision,[164] decreeing that "unrestricted exercise of the right to care implies a *duty of solidarity* among all states of the world".[165] Notably, it deemed the prevailing IHL protections inadequate (especially for medical volunteers not working for the ICRC or a State), and advocated that a UN 'charter' for medical-humanitarian

[159] See, for example, UNSC Res 1296 (19 April 2000) UN Doc S/RES/1296; UNSC Res 2139 (22 February 2014) UN Doc S/RES/2139; UNSC Res 2127 (5 December 2013) UN Doc S/RES/2127.

[160] Jean-Jules Fiset, 'Les privilèges et immunités humanitaires' (1997) 38(1) Les Cahiers de droit 119, 135; Beigbeder (n 2) 347.

[161] Beigbeder (n 2) 347; Fiset (n 160) 135.

[162] Beigbeder (n 2) 348.

[163] ibid.

[164] Council of Europe Resolution 904 (1988) "on the protection of humanitarian medical missions", para 2.

[165] ibid para 3 (emphasis added).

missions be "given the same universal recognition".[166] In short, the rights and responsibilities suggested for relevant personnel were largely similar to those being demanded by no-borders MNGOs at that time.[167]

Beigbeder describes the Council of Europe Resolution as "a constructive compromise between the traditional values and practices of the Red Cross, and the more activist and impatient demands of the 'no-border' Movement leaders".[168] The document, however, did not answer two key questions. The first regards the nature of an international body which could ascertain MNGOs' fulfilment of the requisite principles of humanitarianism, impartiality and neutrality.[169] Beigbeder suggests creating a "specific body of international, independent health-specialists for this purpose… in close consultation with the ICRC".[170] This would serve to allay fears of undue political influence being wielded by wealthy, powerful States. The second question regards the rights of access to a warring State's territory. The Resolution does not mention issues surrounding consent to international assistance, nor whether the duty of solidarity[171] imposed on States would necessitate unimpeded access to MNGOs fulfilling internationally established criteria.

B. Proposed privileges and immunities for medical-humanitarian personnel

As the law stands, IHL protection does not accord personal privileges to medical-humanitarian personnel but is a by-product of protection guaranteed towards vulnerable civilians.[172] Fiset suggests, therefore, that an international convention be drafted specifically to grant MNGOs certain legal privileges and immunities. The latter idea is not novel insofar as it concerns humanitarian personnel. In 1971, for example, the UN General Assembly called for governments of States receiving humanitarian assistance "[t]o consider appropriate legislative

[166] ibid para 11.
[167] ibid Appendix.
[168] Beigbeder (n 2) 351.
[169] ibid.
[170] ibid.
[171] Council of Europe Resolution (n 164) para 11.
[172] Mackintosh (n 6) 117–118.

or other measures to facilitate the receipt of aid, including… necessary privileges and immunities for relief units".[173]

Fiset contends that, in emergency situations, medical-humanitarians should have clear-cut, *functional* privileges[174] such as a 'right' of entry to a State whose population requires emergency assistance, a right to help victims, jurisdictional immunity, and guarantees of protection from attack.[175] These would be in line with the concept of '*l'intérêt de la fonction*',[176] through which diplomats[177] and UN personnel[178] enjoy immunities and privileges requisite to their function.

The catalyst for Fiset's proposition is MNGOs' apparent lack of international status relative to their operational needs.[179] He suggests that such a status be established to safeguard the "interests of humanity".[180] However, international status cannot attach to medical-humanitarian personnel *in abstracto*. Their function in conflict zones derives from the mission itself—a mission mandated by a private MNGO which does not currently have international status equivalent to the UN or sovereign States. Mindful of this, Fiset posits the option of legal personality for relevant organisations.[181] He does not, however, elaborate on how the requisite personality might be negotiated. This question is further investigated below.

C. Could medical-humanitarian NGOs have international legal personality?

An "entity" with international personality has "legal rights and/or obligations and legal capacities directly conferred on it under international law".[182] International law emanates "from state will",[183] with States remaining its primary

[173] UNGA Res 2816, 'Assistance in Cases of Natural Disaster and Other Disaster Situations' (6 December 1971) UN Doc A/RES/2816.
[174] Fiset (n 160) 145–146.
[175] ibid 151, 156–160.
[176] ibid 144–145.
[177] See Vienna Convention on Diplomatic Relations (adopted 14 April 1961, entered into force 24 April 1964) 500 UNTS 95.
[178] Charter of the United Nations (adopted 26 June 1945, entered into force 24 October 1945) 1 UNTS XVI Article 105.
[179] Fiset (n 160) 122.
[180] ibid 146–147.
[181] ibid 151.
[182] Christine Bakker and Luisa Vierucci, 'Introduction: a normative or pragmatic definition of NGOs?' in Dupuy and Vierruci (eds) *NGOs in International Law: Efficiency in Flexibility?* (Edward Elgar 2008) 1.
[183] Roland Portmann, *Legal Personality in International Law* (CUP 2010) 83.

subjects.[184] In recent decades, it has been contended that international personality can derive from States' explicit or implicit recognition.[185] Consistent with this, the ICJ has observed that "subjects of law in any legal system are not necessarily identical in their nature or in the extent of their rights".[186]

Dominicé harnesses this 'recognition conception' to argue that the ICRC has international legal personality.[187] The ICTY confirmed the ICRC's capacity to hold international rights and obligations,[188] noting it had "functions and tasks… directly derived" from the Geneva Conventions (its mandate) ratified by 188 States (its 'recognition').[189] The ICRC is unique: it is a private legal association referred to by jurists as a 'hybrid' or *sui generis* organisation that is neither NGO nor intergovernmental organisation, since its mandate stems from international law.[190]

Whether NGOs can have international legal personality is another question altogether.[191] There may be no academic consensus about their eligibility as potential international subjects,[192] but in "fields of international concern"—once dominated by States—NGOs are increasingly active, no matter the "limited legal regulation of such participation".[193] It is therefore submitted that, by elaborating some formal status for MNGOs, States could more easily "require them to comply with certain international standards",[194] thereby reducing the likelihood of accusations of partiality or of helping 'the enemy'.

Some jurists argue that the international community should adopt a "more flexible recognition of the role played by NGOs in the international legal order… without attempting to place them in a fixed legal framework".[195] This would perhaps be more appropriate for MNGOs such as MSF, which perceive the legal constrictions placed upon the ICRC in humanitarian crises as burdensome. The

[184] Barrat (n 29) 195.
[185] Portmann (n 183) 83–84.
[186] *Advisory Opinion Concerning Reparations for injuries suffered in the service of the United Nations* [1949] ICJ Rep 174 [178].
[187] Christian Dominicé, 'La personnalité juridique internationale du CICR' in Christophe Swinarski (ed) *Studies and essays on international humanitarian law and Red Cross principles* (Martinus Nijhoff 1984) 668–672.
[188] ICTY, *Prosecutor v Simic et al.* (Trial Chamber) Decision on the Prosecution Motion under Rule 73 for a Ruling concerning the Testimony of a Witness, IT-95-9, 27 July 1999 [46].
[189] ibid.
[190] Barrat (n 29) 198–199; Bothe (n 102) 95.
[191] Barrat (n 29) 207–208.
[192] ibid 207.
[193] Bakker and Vierucci (n 182) 6.
[194] Sassòli (n 11) 63.
[195] Bakker and Vierucci (n 182) 6.

ICRC, as "promoter and guardian" of the Geneva Conventions,[196] operates "without exception… with the consent of the parties to the conflict", invoking "legal mandate when reminding them of their obligations",[197] rather than entering territory unauthorised and risking accusations of breached neutrality.

Under the 'flexible' form of international status, NGOs could be granted rights and responsibilities "on a case-by-case basis" if such an approach were 'functional' to the pursued objective.[198] To ascertain an NGO's functional sufficiency, Thuerer advocates recourse to the legal maxim, *ubi societas, ibi ius*[199]—'wherever there is society, there is law'.[200] From this mutual dependence of 'society' and 'law' in international law comes an international society of States encapsulating "the facts of international life".[201] Whilst Dominicé suggests that the ICRC's international "legal consecration" reflects its moral authority as "servant of the suffering",[202] Sandoz has a more pragmatic explanation: the ICRC forced open the doors of the international legal system because it corresponded to international society's *needs at that time*.[203]

This begs the question: does the medical-humanitarian role of 'no-borders' MNGOs in armed conflict correspond with *contemporary* international society's needs, such that States should afford them rights and responsibilities to succour the most vulnerable? As already seen, States jealously guard their territorial sovereignty in times of war. And as "the public sphere has [long] been represented entirely by

196 Rotem Giladi, 'The utility and limits of legal mandate: humanitarian assistance, the International Committee of the Red Cross and mandate ambiguity' in Andrej Zwitter, Christopher Lamont, Hans-Joachim Heintze and Joost Herman (eds) *Humanitarian Action: Global, Regional and Domestic Legal Responses* (CUP 2015) 83.

197 ibid 99.

198 Bakker and Vierucci (n 182) 6.

199 Daniel Thuerer, 'The emergence of non-governmental organizations and transnational enterprises in international law and the changing role of the State' in Rainer Hofmann (ed) *Non-State Actors as New Subjects of International Law* (Duncker & Humblot 1999) 91, cited in Bakker and Vierucci (n 182) 7.

200 Aaron X Fellmeth and Maurice Horwitz, Guide to Latin in International Law (OUP 2009).

201 Jesse Reeves, 'International society and international law' (1921) 15(3) AJIL 361, 368.

202 Dominicé (n 187) 673.

203 Yves Sandoz, 'Le droit d'initiative du Comité international de la Croix-Rouge' (1979) 22 German Yearbook of International Law 352, 371.

the state",[204] Allott observes that international society is a fundamentally "unsocial world".[205]

Nevertheless, Westphalian conceptions of sovereignty and non-intervention[206] are coming under increasing scrutiny.[207] Legal equality among sovereign States holds firm,[208] but there is momentum for "increasing balance… between the rights of sovereign states and the rights of the people who make up their populations".[209] NGOs' role in 'international *civil* society'—a phenomenon, according to Cullen and Morrow, evidencing "the socialisation of international law"[210]—further shapes these 'facts of international life'. What emerges is an increasingly moral, *meta-juridical* basis for MNGOs' arguments when urging States to compromise for the sake of victims' human rights.[211] And, as argued by Judge Ammoun of the ICJ, there is a risk that international law, "in rejecting the moral, social and political elements, described as meta-juridical, [will] become isolated from international realities and their progressive institutions: *ubi societas, ibi ius*".[212]

It is therefore submitted that Thuerer's proposed functional framework would evidence MNGOs' eligibility for international legal status. Just as international law accommodated the already-existent ICRC after that organisation's utility became clear, so too could States recognise MNGOs' unique position in the humanitarian arena, especially given their willingness to tend to civilians who remain inaccessible to State or UN aid agencies. Such a legal development would complement both Fiset's approach to granting legal immunities and privileges, and the "general

[204] Holly Cullen and Karen Morrow, 'International civil society in international law: The growth of NGO participation' (2001) 1 Non-State Actors and International Law 7, 8.

[205] Philip Allott, 'International Law and International Revolution: Reconceiving the World' (Josephine Onoh Memorial Lecture 1989 Hull: Hull University Press, 1989) 8, cited in Cullen and Morrow (ibid) 8.

[206] As enshrined in UN Charter (n 178) Articles 2(1) and 2(7).

[207] Alpaslan Özerdem, 'The "responsibility to protect" in natural disasters: Another excuse for interventionism?' (2010) 10(5) Conflict, Security and Development 693, 700.

[208] Anne Peters, 'Humanity as the A and Ω of sovereignty' (2009) 20(3) EJIL 513, 517.

[209] Özerdem (n 207) 700.

[210] Cullen and Morrow (n 204) 10.

[211] Dupuy (n 53) 31.

[212] *North Sea* (n 58), Separate Opinion of Judge Ammoun.

rationale of IHL to provide protection to categories of persons on the basis of their specific status or function".[213]

D. Towards a right of intervention for mngos' personnel?

Almost three decades ago, Kalshoven and van Reesema posited that better protection for MNGO workers could be secured by coupling victims' "right to receive medical assistance" with recognition of MNGO workers' "right to intervene on humanitarian grounds".[214] They anticipated a groundswell of "[i]nternational pressure… as relief workers are seen to be expelled or imprisoned".[215] As established, however, little progress has been made. Whether war victims have a *positive right* under international law to receive humanitarian assistance remains a moot point.[216] And humanitarian workers seeking a right to intervene encounter a critical stumbling block: Article 3 of the second Additional Protocol enshrines a *general prohibition* on violation of States' sovereignty and non-intervention during NIACs[217]—a prohibition which the Commentary makes clear is also aimed at NGOs.[218]

Yet, there has been progress with regard to the 'humanisation' of IHL. Courts and scholars state that IHL and international human rights law (IHRL) apply in tandem;[219] and the ICJ has confirmed the continued applicability of IHRL in conflict settings.[220] Generally speaking, where human rights law "emphasises granting positive rights to the individual", IHL "protects the interests of individuals through *other means* than the granting of rights".[221] Individuals do not have the capacity to have their IHL rights *enforced*,[222] but this does not preclude them from *claiming* the right in situations of armed conflict.[223] And, in situations where IHL

[213] Breitegger (n 80) 91.

[214] Frits Kalshoven and Charlotte Siewertsz van Reesema, 'Summary of discussions' in Frits Kalshoven (ed) *Assisting the Victims of Armed Conflict and Other Disasters* (Martinus Nijhoff 1989) 205.

[215] ibid.

[216] Dinstein (n 158) 77.

[217] Torrelli (n 100) 237.

[218] Commentary to Additional Protocols (n 112) para 4503. Note, also, that the ICRC/impartial humanitarian organisations' offer of services 'cannot be considered a hostile act' (see Commentary (n 112) para 4505; common Article 3 of the four Geneva Conventions (n 4); AP/II (n 4) Article 1(1).

[219] Barrat (n 29) 18–19.

[220] The ICJ declared IHL to be *lex specialis*, meaning that its norms 'prevail' over the *lex generalis* of IHRL, when appropriate; see *Nuclear Weapons* (n 66) [25]; Dinstein (n 58) 32.

[221] Barrat (n 29) 215 (emphasis added).

[222] ibid.

[223] Hersch Lauterpacht, 'The Subjects of the Law of Nations' (1947) 63 Law Quarterly Review 455, cited in Barrat (n 29) 228.

treaties (particularly relating to IACs) prove "ineffectual", Barrat suggests that human rights law can be used to "clarify IHL guarantees where uncertainty exists".[224]

Consistent with these IHL/IHRL developments, it is submitted that victims of war do have a positive right to receive medical-humanitarian assistance when the State fails to provide it. The ICRC's study on customary IHL indicates civilians' entitlement "to receive humanitarian relief essential to [their] survival",[225] and the right "to make application to… any organization that might assist them".[226] As these are IHL rights, their applicability is confirmed by the international human rights to life, freedom from degrading treatment, and health.[227]

Where does that leave the posited right to intervene on humanitarian grounds?[228] As promulgated by MSF, the right is based on notions about the *universality* of doctors' mission and medicine's transcendence of all borders.[229] If consent is "the expression of sovereignty",[230] then MSF's meta-juridical arguments challenge sovereignty's habitual "precedence over humanity" during conflicts.[231] In this way, it might justify its defiance of States' will, without sacrificing neutrality or impartiality.[232]

For the moment, Ryngaert argues that theories about "'humanising' tendencies in international law" are not universally supported, thus weakening any "claim that a norm limiting the role of state consent has already acquired customary law status".[233] Nevertheless, it is suggested in this article that State sovereignty need not be considered an obstacle to MNGOs' desired right to intervene for their personnel. With the principle of non-intervention considered "as the corollary of… state sovereignty",[234] Peters argues that it is "ultimately grounded in the well-being

[224] Barrat (n 29) 20.
[225] ICRC, 'Customary IHL', Rule 55 (n 18).
[226] ibid; GC/IV (n 4) Article 30(1).
[227] See ICCPR Articles 6 and 7; International Covenant on Economic, Social and Cultural Rights (ICESCR), adopted 16 December 1966, entered into force 3 January 1976) Article 12. See also Barber (n 9) 392–395.
[228] Note that this article does not examine concepts of States' *military* humanitarian intervention or Responsibility to Protect. MSF has criticised such practices for their blurring of humanitarianism and militarism; see Fabrice Weissman, 'Not in our name: Why MSF does not support the "Responsibility to Protect"' (MSF, 3 October 2010) <https://www.msf-crash.org/sites/default/files/2017-05/2749-fw-2010not-in-our-name-msf-and-the-r2p-eng.pdf> accessed 22 July 2018.
[229] Torrelli (n 1) 592.
[230] Torrelli (n 100) 232.
[231] ibid 235.
[232] Chandler (n 39) 683.
[233] Ryngaert (n 13) 13–14.
[234] Peters (n 208) 533.

of natural persons. Non-intervention protects, first, the inhabitants of potential victim-states… and… secures international stability, including the stability of state boundaries."[235]

It seems not unreasonable to extrapolate from this that a State which arbitrarily withholds consent for impartial medical missions cannot cite the principle of non-intervention as a barrier to MNGOs' legitimate entry to the territory. State sovereignty, which Peters likens to "state autonomy", is both a fact of the international legal order and a principle which MNGO personnel, as private actors, cannot possibly undermine.

VI. Conclusion

Medical-humanitarians committed to borderless healthcare have evolved from 'modern-day adventurers'[236] to become a mainstay of victims of conflict in the international community. MNGOs, and MSF in particular, have a standing that affords them the same protection in IHL as that granted to the ICRC. However, there is a large gap to be bridged between that legal protection and the realities of modern conflict. Inter-State wars have been rapidly eclipsed in number by those involving non-State groups, and IHL Conventions and norms arguably have not caught up. In zones inaccessible to external relief, the human toll becomes literally incalculable as the UN is subject to competing political interests, and the ICRC—the exemplar of humanitarianism—stands by traditional interpretations of impartiality and neutrality.

Into this breach step MNGOs such as MSF. 'No-borders' MNGOs, by definition, will man and finance missions in defiance of unpredictable States which arbitrarily withhold consent to entry. Holding themselves to a higher order—medical humanitarianism—comes at a price: unauthorised missions' presence in conflict zones renders legal protection forfeit or subject to the whim of warring parties. This article has examined the prevailing (and relatively nominal) IHL protection for those missions, especially in the context of non-international armed conflict. It has assessed the limits which the current IHL regime places on MNGOs' pursuit of unconditional protection in war zones and analysed alternative legal avenues which might confer enhanced international status.

For too long, the uncomprehensive and, at times, unpredictable nature of IHL protection has left 'no-borders' MNGOs at the mercy of warring parties' caprice. Rather than wait for that legal regime to adapt, MNGOs should receive the legal protection they have requested for decades. The international community may remain reticent to accord international legal personality to NGOs which are

[235] ibid 534.

[236] Torrelli (n 1) 600.

highly active in areas traditionally of State concern, but this should not prohibit an organisation like MSF from being accorded certain rights and responsibilities which complement its role in succouring the sick and wounded. For international law to remain in step with international reality, MNGOs could receive a form of flexible legal status which confirms both their functional necessity and continuing *private* nature, so as to guarantee States' sovereignty. In this context, it is not impossible for international law to strike a balance between the concerns of States and the urgent, medical needs of civilians in wartime.

The Applicability to Dispute Settlement of Most Favoured Nation Clauses in International Investment Agreements

Alperen Afşin Gözlügöl*

I. Introduction

Most Favoured Nation (MFN) clauses are one of the most conventional clauses generally found in international agreements, particularly in trade and investment agreements.[1] In crude terms, the MFN clause prohibits any discrimination against the investor of the contracting state by according any more favourable treatment to the investors of another state. The scope of these clauses may well change according

* B.A. (Bilkent University), LL.M. (Cantab). I would like to thank the anonymous reviewers and editors of the *Cambridge Law Review* (Volume III).

[1] Stanley K Hornbeck, 'The Most-Favoured-Nation-Clause' (1909) 3 AJIL 395; Cristopher Greenwood, 'Reflections on "Most Favoured Nation" Clauses in Bilateral Investment Treaties' in David Caron, Stephan W Shill, Abby Cohen Smutny, and Epaminontas E Triantafilou (eds), *Practicing Virtue* (OUP 2015) 557. For a comprehensive study on the MFN clauses generally, see International Law Commission (ILC), *Final Report of the Study Group on the Most-Favoured-Nation Clause* (29 May 2015) UN Doc A/45/10 <http://legal.un.org/ilc/reports/2015/english/annex.pdf> accessed 4 February 2018.

to their wording,[2] and their applicability in a particular situation depends on the treatment accorded to investors of other nationalities.

These clauses, which seem benign at first blush, have led to one of the greatest divergences in investor-state arbitration in respect of their effect on jurisdiction. The central question is whether, by virtue of the MFN clause in the underlying treaty, an investor may benefit from more favourable dispute settlement provisions found in other treaties concluded by the host state.[3] This question assumes different shapes in different contexts. The underlying treaty, the MFN clause of which the investor invokes, may contain no dispute settlement mechanism, and the investor may be trying to invoke the dispute settlement mechanism found in other treaties through the MFN clause. Alternatively, the underlying treaty may include a dispute resolution mechanism, but it may be limited in its scope and applicability. For example, other treaties may offer a menu of venues for resolving the dispute, while the underlying treaty offers only one forum that may be less favourable from the point of view of the investor. Should the investor have recourse to other venues offered in other treaties through the MFN clause in the underlying treaty? In another scenario, there may be some preliminary conditions which need to be satisfied—the observance of a cooling period or other alternative dispute resolution methods such as conciliation or the exhaustion of local remedies—before the investor is allowed to initiate proceedings before an international tribunal, whereas there are no such conditions in other treaties concluded by the host state. May the investor sidestep these conditions by invoking the MFN clause of the underlying treaty? Moreover, the underlying treaty may provide for the resolution of disputes regarding only the determination of the amount of compensation for expropriation, whereas other

[2] For example, Article 3(2) of the Model Agreement between the Government of the United Kingdom of Great Britain and Northern Ireland and the Government of X for the Promotion and Protection of Investments (1991) ("UK Model BIT") <http://investmentpolicyhub.unctad.org/Download/TreatyFile/2847> accessed 10 January 2018 reads as follows (emphasis added):

> Neither Contracting Party shall in its territory subject nationals or companies of the other Contracting Party, *as regards their management, maintenance, use, enjoyment or disposal of their investments*, to treatment less favourable than that which it accords to its own nationals or companies or to nationals or companies of any third State.

A broader MFN clause can be found in Article 4.2 of the Agreement between the Argentine Republic and the Kingdom of Spain on the Reciprocal Promotion and Protection of Investments (3 October 1991) ("Argentina-Spain BIT") <http://www.wipo.int/wipolex/en/other_treaties/text.jsp?file_id=247353> accessed 10 January 2018: *"[i]n all matters governed by this Agreement*, such treatment shall be no less favourable than that accorded by each Party to investments made in its territory by investors of a third country" (emphasis added).

Greenwood argues that even identically or similarly worded clauses may take different meanings in the light of their context, object, purpose and negotiating history: see Greenwood (n 1) 558.

[3] The term 'underlying treaty' is used throughout in this study to mean the treaty concluded between the host state and the home state of the investor who invokes the MFN clause of this treaty.

treaties offer the dispute settlement mechanism for a broader range of disputes. Should the investor bring its claims that do not relate to matters referred to in the dispute settlement provision of the underlying treaty, but which are covered by the relevant provisions of other treaties of the host state, by relying on the MFN clause of the underlying treaty? These various manifestations of the central problem can be reduced to one question: whether an international tribunal can claim jurisdiction by reference to dispute settlement provisions of another treaty by virtue of the MFN clause in the underlying treaty.[4]

To answer this question thoroughly, one first needs to address the function of the MFN clause. If the function of the MFN clause is to incorporate the more favourable provisions of another treaty into the underlying treaty, then it is reasonable to conclude that more favourable dispute settlement provisions will also be incorporated into the underlying treaty ('incorporation by reference' function—a *renvoi*). An international tribunal can therefore assert jurisdiction under the incorporated provisions of the other treaty, which would not have normally existed under the underlying treaty. On the other hand, if the MFN clause only obliges the contracting state to accord to the nationals of the other contracting state any favourable treatment accorded to nationals of another state, rather than incorporating the provisions of another treaty automatically, the question becomes whether issues relating to dispute settlement—particularly the jurisdiction of tribunals—are 'treatment' in the sense that is used in the MFN clauses; and, if so, how the MFN clause may function reasonably in this case.

Before examining the function of the MFN clauses, one of the fundamental issues that should be determined by a tribunal is whether the MFN clause can be invoked in the first place[5]. This entails, *inter alia*, the *ratione materiae*, *ratione persone*, and *ratione temporis* aspects of the applicability of the treaty. In these respects, if the treaty is not applicable at all, the investor cannot invoke the MFN clause of this treaty before an international tribunal to benefit from other treaties that may be applicable to the investor in these respects. For example, if the investor's operation does not count as an 'investment' as defined under the underlying treaty, the investor cannot invoke the MFN clause of this treaty to benefit from a broader definition

[4] It is admitted that not every condition in the dispute settlement mechanism may relate to the jurisdiction of the international tribunal. For example, it is unclear whether the observance of preliminary conditions such as cooling period concerns the jurisdiction or the admissibility of the claim, or neither of them. However, in this article, for the ease of exposition, all such matters are brought under the 'umbrella' of jurisdiction. This does not affect the validity of the arguments about the applicability of the MFN clauses to the dispute settlement provisions put forward in this article (in Part IV).

[5] See, for example, the *Preliminary Objection in Anglo-Iranian Oil Company Case (United Kingdom v Iran)* [1952] ICJ Rep 93, 109.

of investment found in other bilateral investment treaties (BITs) concluded by the host state which also covers its operation, the reason being that the underlying treaty does not apply to the investor in the first place. It may be asked whether, in the case of jurisdiction, the same reasoning applies. The answer seems to be in the negative.[6] When the MFN clause is resorted to invoke the jurisdiction of a tribunal which would not have existed under the underlying treaty, it cannot be said that the underlying treaty is not applicable at all in the first place. It is applicable; the question is whether the tribunal which purports to apply it can have the jurisdiction to do so through the MFN clause of this treaty. To address this question, the true function of the MFN clauses must first be discussed.

II. The Function Of Most Favoured Nation Clauses

The means to deduce the true function of MFN clauses can be found in an early case in another context: the *Case concerning Rights of Nationals of the United States of America in Morocco*.[7] The dispute related to the consular jurisdiction of the United States in the French Zone of Morocco, which was held to be acquired through the MFN clause of the Treaty between the United States and the Shereefian Empire by relying on the provisions of the Treaties concluded by Morocco with Great Britain and Spain.[8] The International Court of Justice (ICJ) provided the following explanation of how an MFN clause works:

> When the most extensive privileges as regards consular jurisdiction were granted by Morocco to Great Britain in 1856 and to Spain in 1861, these *enured automatically and immediately to the benefit of the other Powers by virtue of the operation of the most-favoured-nation clauses*.[9]

The ICJ has also ruled that these benefits to the United States were terminated by the surrender by Great Britain in 1937 of its rights of jurisdiction in the French Zone, elaborating further on the function of the MFN clause:

> *The… consideration [of the US] was based on the view that the most-favoured-nation clauses in treaties made with countries like Morocco should be regarded as a form of drafting by reference rather than as a method for the establishment and maintenance of equality of treatment without discrimination amongst the various*

[6] *cf. Daimler AG v Argentine Republic*, ICSID Case No. ARB/05/1, Award (22 August 2012) [199]–[204] <https://www.italaw.com/sites/default/files/case-documents/ita1082.pdf> accessed 30 August 2018.

[7] *Case concerning Rights of Nationals of the United States of America in Morocco (France v United States of America)* [1952] ICJ Rep 176.

[8] ibid 190.

[9] ibid 187 (emphasis added).

> *countries concerned*. According to this view, rights or privileges which a country was entitled to invoke by virtue of a most-favoured-nation clause, and which were in existence at the date of its coming into force, would be incorporated permanently by reference and enjoyed and exercised even after the abrogation of the treaty provisions from which they had been derived.… *[T]his contention is inconsistent* with the intentions of the parties to the treaties now in question. This is shown both by the wording of the particular treaties, and by the general pattern which emerges from an examination of the treaties made by Morocco…. These treaties show that *the intention of the most-favoured-nation clauses was to establish and to maintain at all times fundamental equality without discrimination among all of the countries concerned*.[10]

This clearly indicates that the MFN clause does not serve the purpose of incorporating the more favourable provisions of other treaties concluded by the host country into the underlying treaty.[11] If it did, the further existence and validity of the other treaty would be irrelevant for the investor to obtain more favourable treatment contained in that treaty. When the more favourable treatment comes to an end for other investors of different nationality, however, the investor can no longer rely on the MFN clause.[12] Rather than incorporate the more favourable provisions, the MFN clause—in the words of the ICJ—establishes and maintains equality of treatment. It "operates to secure more favourable treatment for the claiming party".[13] The ICJ, similarly, expresses this effect as benefits "enur[ing] automatically and immediately" to the other party.[14] In other words, when the host state has granted a more favourable treatment to the investors of other states whether it is through a BIT, domestic legislation, or *de facto* practice, the investor of

[10] ibid 191–192 (emphasis added).

[11] Although, in its judgment, the ICJ was prudent enough to confine its comment on the function of the relevant MFN clause to the wording of the particular treaties and general pattern of the treaties made by the host state (which in turn demonstrate the intentions of the contracting parties), there is no reason why the current wording of the MFN clauses in BITs and general pattern of the investment treaties should indicate a difference as regards the function of the MFN clauses. There may, however, be exceptions as indicated below (in Part III).

[12] This point is also emphasised in Article 21 of Final Draft Articles on Most Favoured Nation Clauses by the ILC. See ILC, 'Final Draft Articles on Most Favoured Nation Clauses' (1978) 2 Yearbook of ILC 55.

[13] Zachary Douglas, 'The MFN Clause in Investment Arbitration: Treaty Arbitration Off the Rails' (2001) 2 JIDS 97, 105. *cf.* Stephan W Schill, 'Allocating Adjudicatory Authority: Most-Favoured-Nation Clauses as a Basis of Jurisdiction—A Reply to Zachary Douglas' (2011) 2 JIDS 353.

[14] *Case concerning Rights of Nationals of the United States of America in Morocco (France v United States of America)* (n 7) 187.

the home state is automatically and immediately entitled to these benefits as well. This does not mean that they are automatically and immediately incorporated into the underlying treaty. Rather, without any intermediary act, the host state becomes obliged to accord these benefits to the investors of home state under the MFN clause of the underlying treaty; a failure to do so will result in the breach of the MFN clause, hence the underlying treaty. What the MFN clause does not do, however, is to "rewrite the terms of a treaty".[15]

In summary, the MFN provision is a substantive obligation, which is to be invoked by the investor claiming that there has been a breach of the MFN clause. To be able to succeed in its claim, the investor has to show that the host state has granted more favourable treatment to other investors of a different nationality. If this is shown, it means that the host state also has an obligation to extend the benefits to the investors of the home state under the MFN clause. Failure to do so will result in a breach of the MFN clause.[16] In these cases, an international tribunal may: (a) grant declaratory relief, declaring that the investor is entitled to these benefits, or reversely, that the host state is under an obligation to extend these benefits to the investor; (b) award damages to the investor to compensate for the harm that it has suffered as a result of the deprivation of the benefits to which it is entitled under the MFN clause; or (c) order specific performance if it is possible and proper in the circumstances of the case (although this is rare in the investment context, and sometimes not allowed).[17] This last remedy essentially compels the host state to provide relevant benefits to the investor.

[15] Douglas (n 13) 105; Greenwood (n 1) 560–561.

[16] The MFN clauses can be deemed to contain an ancillary obligation. That is, once the contracting state grants a more favourable treatment to investors of different nationality, it becomes obliged to extend the benefits to the investor of other contracting state. So long as it does so, there is no breach of the MFN clause; however, it is still obliged under it because if it stops doing so, it will breach the MFN clause.

[17] It should be noted that ultimately what remedies are available depends on the relevant BIT. However, most investment treaties do not specifically state the consequences of a failure by the contracting state to comply with its treaty obligations. They, however, usually provide that tribunals are to decide disputes also in accordance with 'international law' (or some variation of that formulation). Consequently, tribunals look to customary international law for remedies for internationally wrongful acts in investment treaty sphere. An authoritative statement of customary international law on this matter can be found in International Law Commission's Articles on Responsibility of States for Internationally Wrongful Acts, in particular Articles 28–39. See ILC, 'Draft Articles on Responsibility of States for Internationally Wrongful Acts, with commentaries' (2001) 2 Yearbook of ILC 31. For more detailed information on the remedies in investor-state arbitration, see Borzu Sabahi, *Compensation and Restitution in Investor-State Arbitration* (OUP 2011) 43–133; Eric De Brabandere, Investment Treaty Arbitration as Public International Law Procedural Aspects and Implications (CUP 2014) 175–201; Jeswald W. Salacuse, *The Law of Investment Treaties* (2nd edn, OUP 2015) 436–452.

III. The Jurisprudence Of Investment Tribunals

Having addressed the proper function of the MFN clause as established by early authorities, the relevant jurisprudence of investment tribunals will be briefly examined. The first case that addressed the issue of applicability of MFN clauses to dispute settlement was *Maffezini v Spain*.[18] The dispute settlement provision of the Argentina-Spain BIT provided for a six-month negotiation period, and the submission of the dispute to the competent courts of the host state for its resolution, failing which, in eighteen months, the dispute could be brought before an international tribunal constituted under the BIT.[19] The claimant sought to benefit from the dispute settlement provision of the Chile-Spain BIT by invoking the MFN clause of the Argentina-Spain BIT, according to which "in all matters governed by this Agreement, such treatment shall be no less favourable than that accorded by each Party to investments made in its territory by investors of a third country".[20] The Chile-Spain BIT did not require the submission of the dispute to the competent courts for a period of eighteen months.[21]

The *Maffezini* tribunal held that dispute settlement mechanisms form a part of the treatment-protection accorded to investors:

> Notwithstanding the fact that the basic treaty containing the clause does not refer expressly to dispute settlement as covered by the most favoured nation clause,… there are good reasons to conclude that today dispute settlement arrangements are inextricably related to the protection of foreign investors, as they are also related to the protection of rights of traders under treaties of commerce… It follows that such arrangements, even if not strictly a part of the material aspect of the trade and investment policy pursued by treaties of commerce and navigation, were essential for the adequate protection of the rights they sought to guarantee.[22]

The tribunal, however, excluded the application of the MFN clause to the dispute settlement mechanism when "public policy considerations" are in play.[23]

[18] *Emilio Agustín Maffezini v The Kingdom of Spain*, ICSID Case No. ARB/97/7, Decision of the Tribunal on Objections to Jurisdiction (25 January 2000) 5 ICSID Rep 387.

[19] Argentina-Spain BIT (n 2) Article X.

[20] ibid Article IV(2).

[21] Acuerdo Entre La Republica De Chile Y El Reino De España Para La Proteccion Y Fomento Reciprocos De Inversiones (signed 2 October 1991, entered into force 28 March 1994) ("Chile-Spain BIT")
<http://investmentpolicyhub.unctad.org/Download/TreatyFile/708> accessed 11 January 2018.

[22] *Maffezini v Spain* (n 18) [54].

[23] ibid [62]. The tribunal defines "public policy considerations" as fundamental conditions that contracting parties might have envisaged for their acceptance of the agreement in question.

This is so, according to the tribunal, in cases where the exhaustion of local remedies is required; a fork-in-the-road clause is stipulated; a particular forum is specified; or the parties have submitted to a highly institutionalised system of arbitration (such as under the North American Free Trade Agreement (NAFTA)).[24]

It is understandable that the tribunal held in *Maffezini* that the dispute settlement mechanism offered to investors of different nationalities constituted different treatment, given the role that venue plays in such disputes and the burdens that investors have to go through before initiating proceedings in investment arbitration. However, what the tribunal meant by "public policy considerations", and how it differentiated those that are not included in this category, is not clear. Firstly, the tribunal defined "public policy considerations" as "fundamental conditions that contracting parties might have envisaged for their acceptance of the agreement in question",[25] but this definition instead connotes specifically agreed conditions in the sense of holding parties to their bargain and not allowing the investor to override these specifically agreed conditions. Secondly, the tribunal did not consider the prior resort to domestic courts (for a period of eighteen months), which was included in the Argentina-Spain BIT but not in the Chile-Spain BIT, as reflecting a public policy consideration.[26] It is doubtful that this requirement is so different in terms of public policy considerations from a heavier requirement of exhaustion of local remedies, which was accepted by the tribunal to be included in these considerations. Put it differently, it is not clear why it was not a fundamental condition for the acceptance of the agreement. Furthermore, the tribunal itself did not devise a test, but rather enumerated a few circumstances that it deemed in connection with public policy, leaving it to other tribunals to identify further similar situations.[27]

In *Plama v Bulgaria*,[28] the claimant sought to rely on the dispute settlement provision of the Bulgaria-Finland BIT, which provided for arbitration by the International Centre for Settlement of Investment Disputes (ICSID) in respect of any dispute.[29] It invoked the MFN clause of the Bulgaria-Cyprus BIT, which, in turn, offered *ad hoc* arbitration only for disputes regarding the amount of

[24] ibid [63].
[25] ibid [62].
[26] ibid [64].
[27] ibid [63].
[28] *Plama Consortium Limited v Republic of Bulgaria*, ICSID Case No. ARB/03/24, Decision on Jurisdiction (8 February 2005) 13 ICSID Rep 268.
[29] Article 8 of the Agreement between the Government of The Republic of Finland and the Government of the Republic of Bulgaria on the Promotion and Protection of Investments (signed 3 October 1997, entered into force 16 April 1999) ("Bulgaria-Finland BIT") <http://investmentpolicyhub.unctad.org/Download/TreatyFile/527> accessed 10 January 2018.

compensation for expropriation.[30] The *Plama* tribunal, unlike the *Maffezini* tribunal, held that it was not clear whether the term "treatment" in the MFN clause includes dispute settlement provisions found in other BITs (although it was considered to be irrelevant to address).[31] The *Plama* tribunal further considered, in relation to incorporation by reference by the MFN clause, that:

> …[A] reference may in and of itself not be sufficient; the reference is required to be such as to make the arbitration clause part of the contract (*i.e.*, in this case, the Bulgaria-Cyprus BIT). The reference must be such that the parties' intention to import the arbitration provision of the other agreement is clear and unambiguous. A clause reading "a treatment which is not less favourable than that accorded to investments by investors of third states"… cannot be said to be a typical incorporation by reference clause as appearing in ordinary [commercial] contracts. It creates doubt whether the reference to the other document (in this case the other BITs concluded by Bulgaria) clearly and unambiguously includes a reference to the dispute settlement provisions contained in those BITs. [32]

The tribunal seems to accept the incorporation by reference function attributed to the MFN clauses; however, it does not accept that the MFN clause is sufficient to incorporate the dispute settlement provisions of other BITs given that, usually, any reference to arbitration clauses contained in other documents must be clear and unambiguous. Even if the incorporation by reference function of the MFN clause is to be accepted, it is difficult to understand why a reference to other BITs by the MFN clause is not a clear and unambiguous reference to the dispute settlement provisions contained within those BITs, as was asserted by the *Plama* tribunal. This contention is equivalent to requiring the parties to refer to each and every provision of another BIT separately while they can refer to other BITs as a whole to indicate the provisions of those BITs.[33]

[30] Article 4 of the Agreement between the Government of The People's Republic of Bulgaria and the Government of the Republic of Cyprus on Mutual Encouragement and Protection of Investments (signed 12 November 1987, entered into force 18 May 1988) ("Bulgaria-Cyprus BIT") <http://investmentpolicyhub.unctad.org/Download/TreatyFile/522> accessed 10 January 2018.

[31] *Plama v Bulgaria* (n 28) [189].

[32] ibid [200].

[33] This reasoning is aptly expressed in legal maxims in Latin: *in toto et pars continetur* or *in eo quod plus sit semper inest et*, which respectively mean that 'the part is also included in the whole' or 'in the greater is always included the lesser'.

A more important reason that led the *Plama* tribunal to reject the applicability of the MFN clause to dispute settlement provisions seems to be the idea that such provisions are "specifically negotiated" by the parties:

> It is also not evident that when parties have agreed in a particular BIT on a specific dispute resolution mechanism, as is the case with the Bulgaria-Cyprus BIT (*ad hoc* arbitration), their agreement to most-favored nation treatment means that they intended that, by operation of the MFN clause, their specific agreement on such a dispute settlement mechanism could be replaced by a totally different dispute resolution mechanism (ICSID arbitration). It is one thing to add to the treatment provided in one treaty more favorable treatment provided elsewhere. It is quite another thing to replace a procedure specifically negotiated by parties with an entirely different mechanism.[34]

> Based on this proposition, it is rather more plausible to suggest that, when there is a specifically agreed dispute settlement mechanism, a general reference to more favourable treatment found in other BITs by the MFN clause does not suffice to replace it. As the *Plama* tribunal put it elsewhere:

> [A]n MFN provision in a basic treaty does not incorporate by reference dispute settlement provisions in whole or in part set forth in another treaty, unless the MFN provision in the basic treaty leaves no doubt that the Contracting Parties intended to incorporate them.[35]

As a result, the tribunal concluded that the MFN clause of the Bulgaria-Cyprus BIT cannot be construed as giving consent to submit a dispute under the Bulgaria-Cyprus BIT to ICSID arbitration.[36]

Nevertheless, it is submitted that the idea of dispute settlement provisions in a BIT being 'specifically agreed' does not provide much help. It is hard to contemplate that other provisions of the BIT are not 'specifically agreed'.[37] There is no spectrum of value attached to the consent of the state to the provisions of a

[34] *Plama v Bulgaria* (n 28) [209].
[35] ibid [223].
[36] ibid [227].
[37] Yas Banifatemi, 'Most Favoured Nation Treatment in Investment Arbitration' in Andrea Bjorklund, Ian Laird, and Sergey Ripinsky (eds), *Investment Treaty Law: Current Issues III* (BIICL 2009). See also the Concurring and Dissenting Opinion of Professor Brigitte Stern in *Impregilo S.p.A. v. Argentine Republic*, ICSID Case No. ARB/07/17, 21 June 2011 [21]–[24] <https://www.italaw.com/sites/default/files/case-documents/ita0420.pdf> accessed 30 August 2018.

treaty in the law of treaties. Even if it is accepted that other provisions are not as 'specifically agreed' as dispute resolution clauses, it is not difficult to find specifically agreed aspects of other provisions (as some include specific exceptions or each clause may have different elements of applicability that may render them specific to each other).[38] In such a case, it is doubtful whether the *Plama* tribunal would find the MFN clause inapplicable because to follow such an approach would be self-defeating, and contrary to the very purpose of the MFN clause.

Since the *Maffezini* and *Plama* cases, tribunals have been divided in terms of the precedent that they follow. In line with the *Maffezini* precedent, various tribunals have held that the respective claimants did not need to have recourse first to the local courts of the host country for a period of eighteen months, considering that dispute settlement mechanisms are a part of protection of investment and the treatment accorded to investors.[39] Likewise, the tribunal in *RosInvestCo v Russia* allowed the claimant to rely upon the MFN clause to benefit from a broader dispute resolution provision instead of a limited procedure concerning the amount or payment of compensation in case of expropriation.[40] Taking the opposite line, following

[38] For example, in *CME v Czech Republic*, the relevant BIT provided for "just compensation" representing the "genuine value of the investment affected" in expropriation cases. The tribunal also relied on the MFN clause of this BIT to determine the compensation on the basis of the "fair market value" as provided for in some other BITs concluded by the respondent (should it is accepted that "just compensation" representing the "genuine value" is less than "fair market value"), see *CME v Czech Republic*, UNCITRAL, Final Award (14 March 2003) 9 ICSID Rep 264 [500]. See further *MTD Equity Sdn Bhd* and *MTD Chile SA v Chile*, ICSID Case No. ARB/01/7, Award (21 May 2004) 12 ICSID Rep 6 [100]–[104] and [197] ff, where the tribunal expanded the application of the fair and equitable treatment obligation in the Chile-Malaysia BIT, referring to the corresponding provisions in Chile-Croatia and Denmark-Chile BITs through the MFN clause.

[39] *Siemens A.G. v The Argentine Republic*, ICSID Case No. ARB/02/8, Decision on Jurisdiction (3 August 2004) 12 ICSID Rep 171 [102]–[103]; *Gas Natural SDG, S.A. v The Argentine Republic*, ICSID Case No. ARB/03/10, Decision of the Tribunal on Preliminary Questions on Jurisdiction (17 June 2005) 14 ICSID Rep 282 [31]; *Suez, Sociedad General de Aguas de Barcelona S.A., and InterAguas Servicios Integrales del Agua S.A. v The Argentine Republic*, ICSID Case No. ARB/03/17, Decision on Jurisdiction (16 May 2006) [55]–[57] <https://www.italaw.com/sites/default/files/case-documents/ita0807.pdf> accessed 10 January 2018; *National Grid plc v The Argentine Republic*, UNCITRAL, Decision on Jurisdiction (20 June 2006) [92]–[93] <https://www.italaw.com/sites/default/files/case-documents/ita0553.pdf> accessed 10 January 2018 (the tribunal, however, here seems to approve the Plama tribunal in that the claimant cannot create consent to ICSID arbitration through the MFN clause when there was none before); *Suez, Sociedad General de Aguas de Barcelona, S.A. and Vivendi Universal, S.A. v Argentine Republic*, ICSID Case No. ARB/03/19, Decision on Jurisdiction (3 August 2006) [55]–[59] < https://www.italaw.com/sites/default/files/case-documents/ita0819.pdf> accessed 4 February 2018.

[40] *RosInvestCo UK Ltd v The Russian Federation*, SCC Case No. V079/2005, Award on Jurisdiction (1 October 2007) <https://www.italaw.com/sites/default/files/case-documents/ita0719.pdf> [130]–[132] accessed 10 January 2008.

the *Plama* precedent,[41] various tribunals have differentiated between substantive protections afforded to investors and dispute settlement provisions, concluding that the MFN clause does not extend to the latter unless clearly expressed.[42]

It is submitted that the practice of distinguishing between substantive protections and procedural matters (*i.e.* dispute settlement provisions) in the application of the MFN clauses is not warranted, given that there is no indication to this effect in the wording of the MFN clauses.[43] The key term here is 'treatment'. States are under an obligation not to accord more favourable treatment to other investors of different nationalities. It cannot be denied that the state treats investors differently when it offers the opportunity of international arbitration to one right away but none of such opportunity for the other, or not unless some preliminary conditions are fulfilled.[44] The answer to the question of whether the latter is less

[41] The first case that refused to follow the *Maffezini v Spain* (n 18) decision was *Salini Costruttori S.p.A. and Italstrade S.p.A. v The Hashemite Kingdom of Jordan*, ICSID Case No. ARB/02/13, Decision on Jurisdiction (29 November 2004) 14 ICSID Rep 303. The subsequent cases, however, refer to *Maffezini* (n 18) and *Plama* (n 28) when following either of the lines.

[42] *Vladimir Berschader and Moïse Berschader v The Russian Federation*, SCC Case No. 080/2004, Award (21 April 2006) [179]–[181] <https://www.italaw.com/sites/default/files/case-documents/ita0079_0.pdf> accessed 10 January 2018; *Telenor Mobile Communications A.S. v The Republic of Hungary*, ICSID Case No. ARB/04/15, Award (13 September 2006) 17 ICSID Rep 170 [92].

[43] Banifatemi (n 37) 269; Schill (n 13) 370. It is also argued that "by reason of the *'effet utile'* the MFN clause always covers the dispute settlement mechanism, unless the opposite intention of the Contracting states can be demonstrated": see Yannick Radi, 'The Application of the Most-Favoured-Nation clause to the Dispute Settlement Provisions of Bilateral Investment Treaties: Domesticating the 'Trojan Horse' (2007) 18 EJIL 757, 757.

[44] Banifatemi (n 37) 270; Schill (n 13) 370.

favourable than the former must not be difficult.[45] The availability of dispute settlement mechanisms is just as important for investors as their substantive rights. This can be understood perhaps if one thinks of the fact that one of the most important factors for the decision of the investor whether to vindicate its substantive rights is the path it has to walk to do so.

Undoubtedly, however, when the MFN clause itself regulates its applicability, it is to be followed. For instance, the MFN provision may limit its applicability to certain areas as seen in the Article 1103(2) of NAFTA where the MFN provision applies "with respect to the establishment, acquisition, expansion, management, conduct, operation, and sale or other disposition of investments".[46] In another instance, the MFN clause may explicitly stipulate that it is applicable with regard to dispute settlement provisions such as Article 3(3) of the Model UK BIT.[47]

Overall, it is submitted that, even if the parties did not clearly express their intention to extend the MFN clauses to dispute settlement provisions, the MFN provisions found in most treaties are capable of covering dispute settlement

[45] Douglas asserts that it amounts to a "value judgment about the relative merits of recourse to domestic courts versus recourse to international tribunals": see Douglas (n 13) 111. He also considers that holding procedural protections equivalent to substantive protections in terms of 'treatment' involves a 'value judgment': see Douglas (n 13) 112. It is true that in determining favourability, the tribunal would make a value judgment, but this concern is equally valid for the application of the MFN clause in other contexts, not just with regard to favourability of fora. The determination of the relative favourability must inevitably be made from the point of view of the investors; however, it is not a subjective assessment, but rather an objective assessment: how a reasonable investor if in the position of the investor in the present case would assess the relative favourability, not that how the investor in the present case assesses it. Another more relevant concern would be the principle of comity. As above said, the application of the MFN clauses to dispute settlement provisions inevitably includes evaluations with regard to juridical system of the states. This may harm the comity, and the respect due to the internal functioning and sovereignty of the states. Therefore, it is advisable that international tribunals refrain from making far-reaching comments on the juridical system of the host states. On the other hand, Paparinskis argues that "the ordinary meaning of 'favourable' would require the relevant matters to be sufficiently comparable to establish the relationship of lesser and greater favourability. Most procedural matters do not seem to be capable of such a relationship, reflecting either very different legal techniques without obvious benchmarks for objective comparison or self-judging ad hoc peculiar conveniences of the particular situation", see Martins Paparinskis, 'MFN Clauses and International Dispute Settlement: Moving beyond *Maffezini* and *Plama*?' (2011) 26 ICSID Review 14, 57–58.

[46] North American Free Trade Agreement (United States–Canada–Mexico) (adopted 12 December 1992, entered into force 1 January 1994) ("NAFTA") 32 ILM 289, <https://www.nafta-sec-alena.org/Home/Texts-of-the-Agreement/North-American-Free-Trade-Agreement?mvid=1&seccid=539c50ef-51c1-489b-808b-9e20c9872d25#A1103> accessed 10 January 2018.

[47] UK Model BIT (n 2) Article 3.3 "for the avoidance of doubt, it is confirmed that the treatment provided for in paragraphs (1) and (2) above [national and most favoured nation treatment] shall apply to the provisions of Articles 1 to 12 of this Agreement" and thus includes dispute settlement provision provided in Article 8.

provisions unless expressly excluded by the contracting parties.[48] In short, despite the various reasons and qualifications that have been put forward to preclude the applicability of the MFN clause to dispute settlement mechanisms, the MFN provisions seem as such to be capable of covering them under 'treatment' that must be assessed in terms of favourability to investor as well. The question then becomes how the MFN clause may function reasonably in such a case, especially when there are certain other obstacles to this applicability.

IV. Obstacles To The Applicability Of Mfn Clauses To Dispute Settlement Provisions

As noted above, the function of the MFN clause is not to incorporate more favourable terms of other BITs to which the host state of the investor is a party into the underlying BIT. Therefore, it does not also replace the existing dispute settlement mechanisms with other more favourable mechanisms found in other BITs. If it did so, there would be no room to discuss the applicability of the MFN clauses to the dispute settlement provisions. The MFN clause in a BIT obliges the host state to extend the benefits that it has granted to other investors through several means (*i.e.* treaty, legislation, regulations, *de facto* practice) to the investor of the home state; if it does not, it will have breached the MFN clause, and hence the underlying treaty. This system works relatively easily in other contexts. For example, suppose that the host state permitted the investors of other states to operate without a licence; however, it required such licence from some other investors operating in the same sector and originating from a specific state that has concluded a BIT containing the MFN clause with the host state. Those investors may request from the host state the same treatment to which they are entitled by virtue of the MFN clause, or reversely, the treatment that the host state is under an obligation to procure. If that is denied, the investors may initiate proceedings, and allege the breach of the MFN clause, a substantive obligation. The tribunal will declare that the investors are entitled to the same benefit and that the host state is obliged to extend this benefit to the claiming investors, and rule upon compensation for the harm suffered, if any. In such a case, the harm may be the cost that the investors have incurred to obtain the licence, and the loss of economic benefits that may be incurred because of any delay that obtaining licence has caused in the operation of the investors. The tribunal may be reluctant to order that the operation of the investors be allowed without any licence as it may consider such an order an undue interference with state's sovereignty or unenforceable.[49] These are, in a nutshell, remedies for the breach of the MFN clause. The important point to note is that when the investor invokes the MFN clause, it must put forward and prove its breach by the state; it is

[48] Banifatemi (n 37) 272–273.

[49] De Brabandere (n 17) 184–185.

a substantive obligation, and for an international tribunal to rule on any remedy, there must be a prior breach of a substantive obligation by the state.[50]

It is submitted that this relatively straightforward application of the MFN provisions is not transferable to the sphere of the dispute settlement provisions, and this must be the reason for the proposition that the MFN clauses do not apply to the dispute settlement provisions.

Let us now consider the situation where the investor tries to invoke the jurisdiction of an international tribunal, which would not have existed under the underlying treaty, through the MFN clause depending on the more favourable dispute settlement provisions of other treaties.[51]

If the tribunal does not assert jurisdiction following the objections of the host state, there will be a breach of the MFN clause, because the contracting state would have failed to accord the more favourable treatment to investors under the underlying treaty that it accords to other investors under other investment treaties.[52] The way that the state may accord the same treatment in such a case is to waive any jurisdictional objections and submit to the jurisdiction of the international tribunal. If it does not, what may the cure be for the breach of the MFN clause then? As noted above, there are a few remedies that a tribunal may grant upon a breach of the MFN clause. However, this situation—the breach of the MFN clause— arises on the assumption that the tribunal declines jurisdiction. How can it rule upon any remedy when it declines jurisdiction and this very act is the cause of the breach of the MFN clause in the first place? Might it be suggested that the investor then resorts to the international tribunal constituted as prescribed by the underlying treaty, and asserts damages for the breach of the MFN clause? It is not reasonable. The whole purpose of the applicability of the MFN clauses to the dispute settlement mechanisms for the investors is that they litigate their dispute in their preferred forum. After all, how can they prove any damages that they may have incurred for not being able to litigate in their preferred forum? Moreover, to which act of the state will the liability be attributed is it objecting to the jurisdiction

[50] Brigitte Stern, 'The Elements of an Internationally Wrongful Act' in James Crawford, Alain Pellet, Simon Olleson, Kate Parlett (Assistant) (eds), *The Law of International Responsibility* (OUP 2010) 193–218, 210; Douglas (n 13) 104.

[51] See also Greenwood (n 1) 561–563.

[52] Of course, this conclusion depends on whether the MFN clause, in particular, the 'treatment' covers dispute settlement; if it does not, the state will not be under any obligation at all, thus cannot breach the MFN clause by failing to offer more favourable dispute settlement to the investor of the home state. However, although this matter is contentious, it was argued above that they normally cover the dispute resolution as well unless clearly excluded by the parties (in Part III).

of the first tribunal in the first place?[53] However, the immediate reason that has led to the breach of the MFN clause is that the international tribunal declined jurisdiction. Objecting to the jurisdiction of the first tribunal is ancillary to it. The only conceivable remedy then seems to be that the international tribunal should assert jurisdiction in the first place. That is, by asserting jurisdiction in the first place and dismissing the objections of the respondent state, the tribunal prevents the state from breaching the MFN clause. In a sense, the award in favour of the investor for the breach of the MFN clause becomes the finding of jurisdiction by the tribunal at the preliminary stage of arbitration without any breach of the MFN provision in fact. As explained in the preceding paragraph, however, invoking the MFN clause strictly means claiming a breach of the MFN provision, a wrongful act of the state that has occurred. Without any breach occurring, the tribunal may not rule on any remedy. The finding of jurisdiction, however, removes any possibility of the MFN clause being breached. In brief, the fact that an international tribunal asserts jurisdiction based on the MFN clause, accepting the claiming investor's contentions, runs counter to how the MFN clause normally functions.

Even if one ignores this fundamental rule, the application of the MFN clause to the dispute settlement provisions is beset with further difficulties. By finding that it has jurisdiction, the tribunal effectively forces the respondent state to accord more favourable treatment to the claimant investor, through making an order of specific performance. This aspect alone already indicates the unfeasibility of the application of the MFN clause to the dispute settlement provisions. There may be instances where an effective order of specific performance is not at the disposal of the tribunal, as is the case under NAFTA.[54] In those cases, it is inevitable that the tribunal cannot find jurisdiction. Moreover, in finding that it has jurisdiction, the tribunal not only orders the specific performance in nature, but also enforces this order itself. While one may question the mischief of the tribunal being able to enforce its award itself, it is worthy of note that it is an enforcement of specific performance at the preliminary stage of very arbitration itself. These peculiarities

[53] Douglas rightly asks how the respondent state may reasonably be expected to waive the jurisdictional objections it has in the face of the cardinal principle of any adjudication, the principle of procedural equality: see Douglas (n 13) 104.

[54] In this system, the tribunal may order only monetary damages or restitution of property in which case the state retains the right to pay monetary damages instead. As can be seen, the only kind of an order similar to specific performance is the restitution of property (not applicable in this context) which is also limited as states can still rectify the breach by monetary damages, see NAFTA Article 1135. Similarly, The Energy Charter Treaty (adopted 17 December 1994) ("ECT") 2080 UNTS 95 Article 26(8) provides that "[a]n award of arbitration… shall provide that the Contracting Party may pay monetary damages in lieu of any other remedy granted", thus leaving the award of specific performance at the discretion of the contracting state. Again, this also raises the question of how the arbitral tribunal finding jurisdiction under the MFN clause, which is an award of specific performance in nature, may specify the quantum of damages that the state may pay in lieu thereof.

also demonstrate the non-applicability of the MFN clause to the dispute settlement mechanisms.

On the whole, these considerations show that the MFN clause cannot function as it normally does when applied with regard to dispute settlement mechanisms. By invoking the MFN clause to benefit from the more favourable dispute settlement provisions, investors do not rely upon a breach of the MFN clause, and they cannot do so as shown by the preceding analyses. At most, they may try to show that the MFN clause is the evidence of the intention of the parties to incorporate dispute settlement provisions of other treaties.[55] This is simply not so, as well-established by the long history of the MFN clauses in international law. This difference was also importantly highlighted by the *Renta* tribunal which distinguished between "asserting a breach of the MFN clause and relying upon the MFN clause as evidence of conferring a more expansive jurisdictional mandate to the tribunal".[56] The tribunal held as follows:

> To be clear: the Claimants are not seeking to establish that Russia breached an obligation under the basic treaty (the Spanish BIT) by failing *explicitly* to grant to Spanish investors the same access to international arbitration as the access the Claimants say is enjoyed by Danish investors. The question is instead simply whether Article 5(2) of the Spanish BIT [the MFN provision] evidences Russia's consent that this tribunal's jurisdiction should have an ambit beyond that of Article 10 [dispute settlement provision].[57]

As explained above, however, it is normally not that parties express their intention to incorporate the more favourable provisions of other treaties into their

[55] Douglas usefully draws an analogy between this approach to the MFN clauses and the reliance on an express term in a commercial contract referring to another document which entails dispute resolution provisions: see Douglas (n 13) 106.

[56] Douglas (n 13) 105.

[57] *Renta 4 S.V.S.A, Ahorro Corporación Emergentes F.I., Ahorro Corporación Eurofondo F.I., Rovime Inversiones SICAV S.A., Quasar de Valors SICAV S.A., Orgor de Valores SICAV S.A., GBI 9000 SICAV S.A. v The Russian Federation*, SCC No. 24/2007, Award on Preliminary Objections (20 March 2009) [83] <https://www.italaw.com/sites/default/files/case-documents/ita0714.pdf> accessed 11 January 2018.

treaty through the MFN clause.[58] Rather, the MFN clause seeks to establish and maintain equality of treatment between investors.

V. Conclusion

The MFN clause is a fundamental type of clause, not just in international investment law, but also in international law more generally, and its proper function is to be understood against this background. This article has argued that these clauses ensure equality of treatment among foreign investors of different nationality, rather than rewriting the text of treaties by reference to more favourable provisions. It is this function of the MFN clauses that must be considered when one answers the question of whether they may apply with regard to dispute settlement provisions in international investment agreements.

The jurisprudence of international investment tribunals seems to be divided into two lines of cases, led by the *Maffezini* and *Plama* decisions. While some tribunals have considered the dispute settlement provisions as part of the protection and treatment accorded to the investors, applying the MFN clause in respect of them but along with some qualifications (such as "public policy considerations"), the others have differentiated the dispute settlement provisions from substantive protections and held that the MFN clause does not apply in this regard. It is submitted that both lines of case law are far from convincing. The first line of case law is not clear about, or consistent in, the qualifications that it adds to the application of the MFN clauses to the dispute settlement provisions, whereas the second line of case law fails to appreciate that dispute settlement provisions are also a part of investment protection and treatment in the sense used in the MFN clauses by maintaining an artificial division. Both, however, premise their analyses on a fundamental flaw: they seem to accept that the function of the MFN clauses is to incorporate by reference the more favourable provisions of other treaties. The central question is not whether the dispute settlement provisions are 'treatment', or 'procedural questions rather than substantive protections', or 'specifically agreed', but rather whether the MFN clauses may function properly as regards the dispute settlement provisions as they function in other contexts.

It has been shown in this article that the application of the MFN clauses with regard to dispute resolution clauses is not feasible. Invoking the MFN clause is not enabling the jurisdiction of the international tribunal; rather, it means showing a breach of the MFN clause and seeking remedies from the international tribunal. Be that as it may, when the MFN clause is invoked in this context, there is no breach

[58] Greenwood rightly argues that it "cannot be ruled out, especially where the language or drafting history of the MFN clause and the investor-state arbitration clause indicate that the parties to the BIT intended that the MFN clause apply so as to accord to investors the same access to arbitration as that offered in other BITs": see Greenwood (n 1) 563.

of the MFN clause when the investor seeks to convince the international tribunal to find jurisdiction. It is only when the international tribunal rejects jurisdiction that a breach of the MFN clause may come into existence. While how this will help the investor is very doubtful, how the state will be held liable is not clear. If the tribunal finds jurisdiction, there is no breach of the MFN clause, although the MFN clause is the provision that the investor relies on. This is no less illogical than that international tribunal declares that there is no breach of fair and equitable treatment and at the same time awards damages when the investor relies on the fair and equitable treatment standard. The MFN clause is a substantive obligation, and the investor cannot rely on substantive obligations without a breach. By finding jurisdiction, the tribunal negates in advance the breach of the MFN clause, which in fact amounts to an order of specific performance enforced by the tribunal itself at an interlocutory stage of the arbitral proceedings. How this will be possible when the order of specific performance is not allowed (if allowed, it is a rare practice anyway) is a question which can be readily answered.

Overall, the focus in the current practice of international investment arbitration is in the wrong place. Treatment, of course, covers how favourably the state chooses to solve its disputes with the investors of different nationalities. When the true function of the MFN clause is revealed, however, it is clear that their application with regard to dispute settlement provisions is not viable. This article will have achieved its purpose if it has shown that the reason why the MFN clauses will not apply to dispute settlement provisions is because they *cannot*.

Dispute Settlement in the World Trade Organisation: Moving Towards an Acknowledgement of Stare Decisis

Jia Ying Lim*

I. Introduction

The Dispute Settlement Body (DSB) has been hailed as a central pillar of the World Trade Organisation's (WTO) success.[1] It has compulsory jurisdiction over WTO members (hereinafter, "Members") and stands as the "core linchpin of the whole international trading system",[2] interpreting and upholding the WTO Agreement.[3] From its inception to 2014, the DSB's Appellate Body (AB) has dealt with no less than one-hundred and twenty-nine appeals; the DSB Dispute Panel (hereinafter, "Panel") has handled even more disputes.[4] Following an increasing stream of litigation, the DSB is steadily developing a substantial body of case law on the interpretation and application of the WTO Agreement.

As a result, the controversial issue of the legal status of Panel and AB reports which have been approved by the DSB (Reports) is gaining in prominence and importance by the year. It is still unclear what the position is of such Reports in WTO law: whether they are in themselves a source of legal authority, as part of

* LL.B. (Singapore Management University) (Candidate). I would like to thank Professor Yang Guohua, whose facilitation of and guidance in seminar discussions provided much invaluable insight into the WTO's dispute settlement system. All errors remaining are my own.

1 Adrian TL Chua, 'Precedent and Principles of WTO Panel Jurisprudence' (1998) 16 Berkeley Journal of International Law 171.

2 John H Jackson, *The World Trading System* (2nd edn, MIT Press 1997) 124.

3 For the rest of this article, a broad understanding of the WTO Agreement will be adopted. Thus, any subsequent reference to the WTO Agreement includes the Marrakesh Agreement, its appendices and all related documents such as Accession Protocols. See Marrakesh Agreement Establishing the World Trade Organization (adopted 15 April 1994, entered into force 1 January 1995) 1867 UNTS 154 (hereinafter, "the Marrakesh Agreement").

4 The World Trade Organisation, 'Dispute Settlement: Statistics' <https://www.wto.org/english/tratop_e/dispu_e/stats_e.htm> accessed 26 December 2017.

the *corpus* of law, akin to the status of judicial decisions in common law; or whether they are merely of subsidiary status as part of the *acquis* of WTO law, per civil law.[5]

If Reports do not enjoy legal precedential status, then the consolidated Panel and AB jurisprudence becomes of diminished value. All the hundreds of pages of effort gone into the writing and editing of each Report is limited to the facts at hand and have no value thereafter; the much vaunted transparency of the DSB is thus rendered of limited use. This is a conclusion that instinctively does not sit well with efficiency considerations. Hence, this article will endeavour to give a reasoned legal analysis as to why the Reports should enjoy precedential effect under a doctrine of *stare decisis*.

II. The Concept Of Stare Decisis

The doctrine of *stare decisis* (SD) is a common law concept that, in brief, means "to abide by, or adhere to, decided cases".[6] Accordingly, if courts in prior judgments have laid down a "principle of law as applicable to a certain state of facts, it will adhere to that principle, and apply it to all future cases, where facts are substantially the same, regardless of whether the parties and property are the same".[7] Its purpose is to give the law a "tensile toughness",[8] imbuing the law with a level of consistency and predictability so as to allow its subjects legitimate expectations on the operation of the law, and to be consistent with the rule of law.

However, the doctrine of SD today is no longer strictly binding in the UK and the US, as archetypes of the world's common law jurisdictions. The common law values consistency, but ultimately, the judge's higher obligation is to "his mistress, the law".[9] This is reflected in the development of vertical and horizontal SD over time. The former is the obedience of a lower court to a higher court in the judicial hierarchy, while the latter describes how a judge is bound by or must respect earlier decisions by another court of the same coordinate level. While vertical SD is still strictly followed, the parameters of horizontal SD have been relaxed, particularly with regard to apex courts. In the UK, a House of Lords Practice Statement recognised that "too rigid adherence to precedent may lead to injustice in a particular case", and as such, "while treating former decisions of this House

5 Wooraboon Luanratana and Alessandro Romano, '*Stare Decisis* in the WTO: Myth, Dream or a Siren's Song?' (2014) 48 Journal of World Trade 773, 777 ff.

6 Bryan A Garner, *Black's Law Dictionary* (10th edn, West Group 2014) 1406.

7 ibid.

8 Neil McCormick and Robert Summers, *Interpreting Precedents: A Comparative Study* (Routledge 1997) 355, 396–397.

9 Carleton Allen, *Law in the Making* (6th edn, OUP 1958) 280.

as *normally binding*, [the House would] depart from a previous decision when it appears right to do so".[10] Similarly, in the US, appellate judges "expressly overrule precedents at least two or three times a year in almost every state".[11]

The chief implication of this change is that the highest court is now able to depart from prior precedent. Hence, the greatest weakness of a strict doctrine of SD has been diminished, as a court will not uphold a legal principle simply "because it was laid down in the time of Henry IV".[12] However, this newfound flexibility is as much a two-edged sword as strict SD; it brings heightened fears of unfettered judicial law-making, as it leaves more power and discretion in the hands of appellate judges.

In this article, where SD is mentioned, it refers to the newer, less strict understanding of the doctrine of precedent. This is the modern incarnation of the doctrine after years of progress: the UK and the US, as originators and champions of the SD doctrine, now abide by this less binding variant of the doctrine.

III. Status of Stare Decisis in the Wto

A. Denial of stare decisis

(i) Legislative instruments

Various WTO legal authorities, including the Dispute Settlement Understanding (DSU) itself,[13] have repeatedly emphasised that the common law doctrine of SD has no place in the DSB. The root of this statement can be traced back to Article IX(2) of the Marrakesh Agreement,[14] which confers upon the Ministerial Conference and the General Council the exclusive authority to adopt interpretations of the WTO Agreement. Given that such exclusive authority was explicitly granted to these bodies but not to the DSB or its AB, it is logical to assume that the adopted AB or Panel Reports do not constitute authoritative interpretations of the WTO Agreement.

There are also other indicators to the effect that the DSB's interpretations are not authoritative, such as Article 3(2) of the DSU. Article 3(2) prohibits the DSB from "add[ing] to or diminish[ing] the rights and obligations provided in the

[10] *Practice Statement (HL: Judicial Precedent)* [1966] 1 WLR 1234, [1966] 3 All ER 77.

[11] Allen (n 9) 404.

[12] *Loschiavo v Port Authority* (1983) 58 NY.2d 1040, 1043.

[13] WTO, 'Understanding on Rules and Procedures Governing the Settlement of Disputes', Annex 2 of the Marrakesh Agreement Establishing the World Trade Organization (adopted 15 April 1994, entered into force 1 January 1995) 1869 UNTS 401 (hereinafter, "Dispute Settlement Understanding").

[14] Marrakesh Agreement (n 3) Article IX(2).

covered agreements" via its "recommendations and rulings". If the Reports were to have precedential effect, a decision today between two parties would impact the parameters of a third party member's rights or duties in future cases.[15]

Further, Articles 3(3) and 3(4), in explaining the function of the DSB, focus on how the DSB ensures the "prompt settlement of the situation" or the "satisfactory settlement of the matter", with no indication of any hope or intention to build a body of jurisprudence from the rulings and recommendations.

(I) Panel and ab pronouncements

The *inter partes* rule in the DSU has been repeatedly affirmed by personnel associated with the WTO. The WTO Legal Affairs Division and the Appellate Body Secretariat jointly affirmed that "the Reports… are not binding precedents for other disputes between the same parties on other matters or different parties on the same matter… there is no rule of SD in WTO dispute settlement according to which previous rulings bind panels and the AB".[16]

This sentiment is echoed in several Panel and AB Reports, even under the former General Agreement on Tariffs and Trade of 1947 (GATT 1947). Two cases involving certain European measures on imports of apples from Chile show this point. In *EEC – Apples (1989)*,[17] Chile complained the European Economic Community (EEC) was fixing the prices of apples, and the EEC countered with the exception in Article XI(2)(c)(i) GATT 1947. The same issue, over the same subject and between the same parties, had been raised in an earlier case, *EEC – Apples (1980)*.[18] However, in *EEC – Apples (1989)* the Panel did not rely on such precedent, but re-examined the issue entirely following a different legal reasoning. Indeed,

[15] Raj Bhala, 'The Myth about *Stare Decisis* and International Trade Law (Part One of a Trilogy)' (1999) 14(4) American University International Law Review 845, 879.

[16] WTO, 'Legal Effect of Panel and Appellate Body Reports and DSB Recommendations and Rulings' in WTO, *Dispute Settlement System Training Module: Chapter 7* <https://www.wto.org/english/tratop_e/dispu_e/disp_settlement_cbt_e/c7s2p1_e.htm> accessed 26 December 2017. However, this is not conclusive as to the legal effect of precedents since this is merely a statement on the WTO website but is not incorporated into any legally binding agreement between the Members, nor is it a pronouncement from the two bodies which have been given explicit authority to interpret the WTO Agreement (including whether or not its DSB's judgments have precedential value).

[17] *EEC – Restrictions on Imports of Dessert Apples* (Complaint by Chile) (1989) GATT L/6491 36S/93.

[18] *EEC – Restrictions on Imports of Apples from Chile* (1980) GATT L/5047 27S/98.

the Panel noted that "did not feel it was legally bound by all the details and legal reasoning of the 1980 Panel report".[19]

Similar statements were made in other Reports, such as in *Japan – Taxes on Alcoholic Beverages*,[20] where the AB stated that Panel Reports are "not binding [to subsequent Panels], except with respect to resolving the particular dispute between the parties to that dispute".[21] In addition, the AB has been known to overturn its past decisions. *China – Raw Materials*[22] and *China – Rare Earths*[23] were independent cases decided less than two years apart, yet *China – Rare Earths* decisively rejected its predecessor. This is a notable deviation from the practice of courts under an SD regime, where precedents, if overturned, are generally done so only after a considerable length of time. While such a move is theoretically possible under the doctrine of SD, the low likelihood of its occurrence in a system abiding by the doctrine of SD makes it more probable than not that the Panel and AB do not consider themselves bound by precedents.

B. De facto practice of stare decisis

Despite the apparent inapplicability of the SD doctrine to the DSU's operations, some have observed that the Panel and AB have adopted a *de facto* practice of SD.[24]

(I) Persuasive authority: panel and ab practice

Panel and AB Reports invariably come attached with a table of cases, which list down past Reports cited by parties or panels as relevant to the case at hand. This shows that, in practice, disputes brought before the Panel and AB are not limited to an *inter partes* effect, as every Report may have precedential effect in future cases.

Other practices of the Panel and AB further evince an awareness that their reasoning and decisions are of value beyond the case at hand: in several cases, such as *EEC – Parts and Components*[25] and *Japan – Restrictions on Imports of Certain Agricultural*

[19] *EEC – Apples* (1989) (n 17) [12.1].

[20] *Japan – Taxes on Alcoholic Beverages* (1996) WTO WT/DS8/AB/R.

[21] ibid 14.

[22] *China – Measures Related to the Exportation of Various Raw Materials* (2012) WTO WT/DS394/AB/R.

[23] *China – Measures Related to the Exportation of Rare Earths, Tungsten and Molybdenum* (2014) WTO WT/DS431/AB/R.

[24] For example, see Bhala (n 15) and Anne Scully-Hill and Hans Mahncke, 'The Emergence of the Doctrine of *Stare Decisis* in the WTO Dispute Settlement System' (2009) 36(2) Legal Issues of Economic Integration 133.

[25] *EEC – Regulations on Imports of Parts and Components* (1990) GATT L/6657 - 37S/132.

Products,[26] the Panel and AB pre-emptively reminded parties that their reasoning would, given the circumstances at hand, only be applicable to the specific matter, which implies that the Panel or AB is aware that it may be used *beyond* the specific matter.

In addition, the Panel and AB have on various occasions been asked to, and agreed to, rule on expired measures which no longer fuel live issues. For example, in the famous *US—Woven Wool* case,[27] part of the dispute centred on the validity of US transitional safeguard measures against Indian wool imports. The measures were withdrawn before the Panel reached a decision, yet, India specifically requested that the Panel continue to finalise and release its Report.[28] The only reason for India to do so would be if it believed that the reasoning and decisions could be of use in the *future*, instead of being limited only to the specific situation and parties at hand.

Therefore, it is not only the adjudicatory bodies which assume that their Reports have precedential value; the Members under their jurisdiction have—via their actions—also indicated that they too share a similar belief.[29] However, these examples only go towards showing that the old Reports are referred to in new judgments and thus enjoy persuasive precedential value, but fall short of evidencing a practice of *de facto* SD. It ought to be recalled at this juncture that SD refers to a practice of *normally binding* vertical and horizontal precedent; not quite invariably binding but also not merely persuasive authority.

(II) Specific Cases

Stronger evidence for the *normally* binding nature of precedent can be found elsewhere in the DSB's operations. In particular, the language of SD recurs in Reports: in *Canada – Periodicals,*[30] the AB distinguished a prior Report on the grounds that the part of the Report cited by the USA constituted only "*obiter dicta*" and was, therefore, not binding.[31] The term "*obiter dicta*" and the related idea of the binding "*ratio decidendi*" are singular to the concept of SD as it is known in common law.

The case that is now frequently cited as establishing de facto SD in the DSB is *US—Stainless Steel*.[32] In that case, the AB pronounced that future Panels are not

[26] *Japan – Restrictions on Imports of Certain Agricultural Products* (1988) GATT L/6253 35S/163 [5.4.1.4].
[27] US – *Measures Affecting the Imports of Woven Wool Shirts and Blouses from India* (1997) WTO WT/DS33/AB/R.
[28] ibid 2.
[29] Chua (n 1) 177–178.
[30] *Canada – Certain Measures Concerning Periodicals* (1997) WT/DS31/AB/R.
[31] ibid 33.
[32] *US – Final Anti-Dumping Measures on Stainless Steel from Mexico* (2008) WT/DS344/AB/R.

permitted to "disregard the legal interpretations and the *ratio decidendi* contained in previous [adopted] Appellate Body Reports".[33] The Report acknowledged that "WTO Members attach significance to reasoning provided in previous Panel and Appellate Body reports… [which are] often cited by parties in support of legal arguments…in subsequent disputes".[34] Hence, "[t]he legal interpretation… becomes part and parcel of the acquis of the WTO settlement system".[35]

Importantly, the AB reasoned that apart from the practice of using cases as persuasive precedent, "[WTO] Members recognised the importance of consistency and stability in the interpretation of their rights and obligations under the covered agreements… [which in turn] is essential to promote 'security and predictability' in the dispute settlement system…"[36] As such, to protect Members' legitimate expectations, and to "[ensure] 'security and predictability' in the dispute settlement system, as contemplated by Article 3(2) of the DSU… absent cogent reasons, an adjudicatory body will resolve the same legal question in the same way in a subsequent case".[37] The language of "cogent reasons" echoes the "normally binding precedent" stand in common law jurisdictions—precedents would by default be followed, save certain exceptions. The case concluded that "the Panel's decision to depart from well-established Appellate Body jurisprudence… has serious implications for the proper functioning of the WTO dispute settlement system".[38]

The "cogent reasons" standard was accepted and a further test was adopted in *US – Countervailing and Antidumping Measures*.[39] The case provided four non-exhaustive situations which would justify departure from an otherwise applicable precedent:[40]

> A multilateral interpretation of a provision of the covered agreements under Article IX(2) of the WTO Agreement that departs from a prior Appellate Body interpretation;
>
> A demonstration that a prior Appellate Body interpretation proved to be unworkable in a particular set of circumstances falling within the scope of the relevant obligation at issue;
>
> A demonstration that the Appellate Body's prior interpretation leads

[33] ibid [158].
[34] ibid [160].
[35] ibid [160].
[36] ibid [161].
[37] ibid [160].
[38] ibid [162].
[39] *US – Countervailing and Antidumping Measures* (2014) WTO/DS449/R.
[40] ibid [7.317].

> to a conflict with another provision of a covered agreement that was not raised before the Appellate Body; and
>
> A demonstration that the Appellate Body's interpretation was based on a factually incorrect premise.

The case confirms the *US – Stainless Steel* ruling that precedent is not merely persuasive, but is—to a certain extent—binding (particularly in the sense of vertical SD). Even if a future adjudicatory board find themselves persuaded by legal arguments that reach a different conclusion, they are unable to stray from prior AB rulings. Hence, Panels and the AB are instructed to render their decisions with strong deference to prior cases, which is in practice adherence to SD.

C. Controversy over stare decisis

Given the wealth of sources insistently reassuring members that the doctrine of SD does not apply in the context of the DSB, the mixed signals sent by Panels and the AB are confusing and unjust to members, particularly if the DSB is in truth prohibited from adopting the doctrine of SD. Until this fundamental issue is settled, it is highly likely that parties to a dispute, when faced with undesirable precedent, will attempt to argue that: (a) the AB's prior decisions do not even have high precedential value; and (b) even if they do, the exact standard for the AB or Panels to stray from precedent (the "cogent reasons" test) is too high.

In both situations, the central issue is that whatever the adjudicator pronounces will not be satisfactorily regarded as final. One of the parties will accuse the adjudicator of spinning both the doctrine of SD and the "cogent reason" test (for the doctrine's application) from thin air. Any decision or guidelines on precedent, however, will not be conclusive because the unhappy Member—and any unhappy future litigants—will simply argue that these requirements are not binding in the context of future rulings because AB or Panel rulings do not have precedential value in the first place. If that is the case, the same issue will arise repeatedly as part of an endless cycle, wasting WTO and DSB resources. Hence, for the dispute resolution mechanism to continue functioning efficiently, a conclusive answer must be reached on this matter.

IV. Arguments on the Existence of Stare Decisis per the WTO Agreement

To end the stalemate, it is crucial to identify whether the WTO Agreement envisions a doctrine of SD. In the absence of a clear statement, the next best

option is to identify whether the WTO Agreement *excludes* the operation of the doctrine. It would be unrealistic to hope to find an explicit, conclusive statement in the Agreement on whether the doctrine of SD is applicable, given that scholars and Members have been arguing over this issue for a decade, and would have already reached a unanimous resolution if the answer could so easily be found. Hence, the focus shall be on proving that the WTO Agreement does not in fact forbid the operation of SD.

The subsequent analysis of the WTO Agreement's relevant provisions will, as far as possible, not involve interpretations or applications found in Reports. This is a logical concession as the question under discussion is whether Reports findings have value beyond their specific factual scenario; in particular, whether Reports have precedential value. Hence, it would be circular reasoning to use Report findings to substantiate the argument.

A. WTO agreement does not forbid stare decisis

Earlier, Article IX(2) of the WTO Agreement was identified as the backbone of the argument that the doctrine of SD has no place in the DSB. This is supported by clauses in the DSU agreement itself, which ostensibly lend to the conclusion that the doctrine of SD cannot apply. However, a closer reading of the relevant provisions shows that the WTO Agreement does not reject the doctrine's operation.

(i) Article 3(2) of the DSU

As noted earlier, Article 3(2) states that "[r]ecommendations and rulings of the DSB cannot add to or diminish the rights and obligations provided in the covered agreements", which has been read to exclude the operation of the SD doctrine. The argument goes that if Reports had precedential effect, the rulings of the DSB would affect the rights and obligations of WTO Members in future disputes.[41]

There is, however, an alternative way to understanding Article 3(2) that does not necessitate the conclusion that SD cannot operate in the DSB. Article 3(2) explains that the WTO's dispute settlement mechanism seeks to provide "security and predictability" regarding the operation of the rules in the WTO Agreement, which is achieved when the DSB clarifies the provisions of the WTO Agreement.[42] Subsequently, when enforcing WTO members' rights and obligations—the

[41] Bhala (n 15) 879.

[42] John H Jackson, 'International Law Status of WTO Dispute Settlement Reports: Obligation to Comply or Option to "Buy Out"?' (2004) 98(1) American Journal of International Law 109, 116.

parameters of which are interpreted by the DSB in the Panel or AB reports—the system preserves members' existing rights and obligations. The goal of providing predictability to WTO members is further achieved when a body of jurisprudence is developed, taking precedential effects that can be relied on by both immediate disputants and other WTO members in future disputes.[43] Thus, on a purposive reading of Article 3(2), in the context of dispute resolution proceedings, the DSB does not alter the rights of parties, but merely *explains* existing rights and obligations under the WTO Agreement. That should be the preferred understanding of the DSB's role.

This is akin to the difference between a *discovery* of a legal principle, which is within the purview of common law judges' duties and powers, as opposed to *creating* law, which judges are technically not supposed to do.[44] The taboo against judicial activism is prevalent even in common law jurisdictions. As such, it would be more palatable to any party or Member of the WTO if the DSB's pronouncements on the WTO Agreement, communicated via Panel and AB Reports, are regarded as mere interpretations of existing law. If that is the meaning of Article 3(2), then Panels and the AB when interpreting the WTO Agreement are merely (legitimately) explaining existing rights and obligations, and these interpretations are, therefore, not precluded from holding precedential value.

This is further supported by the logical inference that Article 3(2) cannot possibly be referring to Panels and AB affecting the rights and obligations of Members simply through ordinary interpretation of the WTO Agreement. If a pertinent question of interpretation arises in a dispute, the Panel or AB must necessarily reach an answer on the matter. Such a pronouncement, because of the litigious nature of the dispute, would in all likelihood be favourable to one party but not to the other. If that were, by itself, to constitute illegitimate interpretation, then the DSB would be wholly powerless because it would not be able to settle disputes at all. Thus, according to Article 31 of the VCLT, which requires that a treaty be interpreted in good faith in light of its object and purpose, Article 3(2) should be interpreted in this suggested manner, which avoids the rendering the DSB's dispute-settlement process ineffectual.

If it is accepted that Panels and the AB take on an explanatory role when interpreting and applying provisions, then these pronouncements are automatically capable of having precedential value. The concept of *discovering* the law means that there is necessarily only one pre-existing, objectively correct understanding of the

[43] ibid.

[44] Zecheriah Chafee Jr, 'Do Judges Make or Discover Law?' (1947) 91(5) Proceedings of the American Philosophical Society 405.

law.[45] Panels, and subsequently the AB acting as a check on the Panel, are stating a truth that will be equally valid in future cases as it is in the case at issue.

The only reason against such interpretations having precedential value would be to argue that the DSB is not the appropriate body to make such a pronouncement. This may be justified under the earlier analysis of Article IX(2) of the WTO Agreement,[46] which confers such binding authority—apparently exclusively—on the Ministerial Conference, the General Council and none other.

However, it is difficult to reconcile a strict reading of Article IX(2) of the WTO Agreement with Article 3(2) of the DSU. If Article IX(2) is interpreted to mean that only the Ministerial Conference and the General Council can adopt authoritative interpretations, then the DSU's panel and AB reports are not authoritative interpretations. Thus, the current practice wherein the DSB announces interpretations—apparently not authoritatively, per Article IX(2)—and subsequently enforces judgment[47] causes parties in a dispute to have their rights re-defined and altered by the DSB, violating Article 3(2) of the DSU. This is because if the DSB is not authoritative, its interpretations could be mistaken and enforcing a mistaken judgment would then constitute derogating from members' rights and obligations under the WTO Agreement, which members should be protected from under Article 3(2).

To avoid such derogation from existing rights and obligations, it appears that the only logical solution left is to demand that either the Ministerial Committee or the entire General Council decisively entertain all questions of interpretation by divining one right understanding of the law. This is, however, an unfeasible proposal given the difficulty of obtaining consensus or at minimum a three-quarter majority and the inefficiency of bothering the MC for individual cases.[48] In addition, such a move would render the valued DSU mechanism obsolete and inefficient. As a result, Members must reconsider the implications of reading Article IX(2) as authority against *stare decisis*.

It would not be inconceivable for the DSB to have that authority. Whereas Article IX(2) indeed neglects to explicitly grant exclusive authority to the DSB, the reason could well be that the composition of the General Council and the DSB are identical. Article IV(3) of the WTO Agreement explains that "[t]he General Council shall convene as appropriate to discharge the responsibilities of the

[45] ibid.

[46] See Part III.A.(i) above.

[47] The pronouncement of decisions by the DSB and its binding effect on parties to the dispute is provided for in Article 17(14) of the Dispute Settlement Understanding (n 13).

[48] Marrakesh Agreement (n 3) Articles IX(1) and IX(2).

Dispute Settlement Body provided for in the Dispute Settlement Understanding". Indeed, the equivalence of the General Council and the DSB is implied in Note 3 to the Marrakesh Agreement that speaks of "[d]ecisions by the General Council when convened as the Dispute Settlement Body…" Hence, although these are two distinct legal bodies, they are in practice composed of the same members and might be regarded as *alter egos* of each other, so that the General Council *is* sometimes the DSB. As such, in agreeing to grant exclusive interpretive authority of the WTO Agreement to the General Council in Article IX(2), it can be extrapolated that Members have also granted authority to the General Council's *alter ego*, the DSB.

Only the DSB may make rulings under the DSU, echoing the authority of the General Council. It has the authority to decide whether to adopt a Panel or AB Report, which are merely recommendations and not binding upon the parties to the dispute.[49] When the DSB approves of a Report, it accepts the interpretations of the WTO Agreement contained therein. By virtue of the DSB's status as the *alter ego* of General Council, the General Council can by extension be seen to have accepted the same interpretations. Given that the General Council's interpretations are authoritative, they become timeless interpretations of the WTO Agreement and are, therefore, normally binding precedent on future cases in which the same interpretative issue arises.

Although, it may be unpersuasive to consider the DSB and General Council to be legal *alter egos*,[50] the identical composition of the two bodies does make it less objectionable for the DSB to hold similarly conclusive interpretive authority. Hence, the simplest way to resolve the glaring inconsistency between Article IX(2) WTO and Article 3(2) DSU, would be to openly acknowledge that the DSB does possess interpretive authority.

(II) Other DSU articles

It was earlier noted that other provisions in Article 3 of the DSU appear to imply that the Panel and AB judgments should be limited *only* to the factual scenario at issue. Articles 3(3) and 3(4) in explaining the duties of the DSB repeatedly focus only on "the matter" at hand, without any reference to the DSB contributing to the creation of a body of jurisprudence. It might therefore reasonably be inferred from

[49] Dispute Settlement Understanding (n 13). See in particular Articles 16 and 17(14) which entrusts the right to adopt a report to the DSB, and Article 11, which terms Panel and AB reports as "recommendations".

[50] The author acknowledges that this is likely to be a controversial argument because of the fundamental principle of separate legal entities.

the silence that the latter is not part of the DSU's mandate. There are, however, two reasons that demonstrate that Articles 3(3) and 3(4) are inconclusive on the matter.

Firstly, it is overthinking to infer that it was not intended that the DSU's Reports lack any precedential value simply from the provisions' silence. Article 3 of the DSU is titled "General Provisions" and deals with the day-to-day functions of the DSU, which are indeed primarily to resolve disputes. This should be contrasted with the preamble of the WTO Agreement, which in its recitals comprehensively list the goals and visions of the WTO as a body facilitating international trade.[51] If there were a similar preamble in the DSU, then it would be more reasonable to expect that the DSU's goals and functions be comprehensively stated, including any long-term goals to build a body of precedent. Hence, in the absence of such an overarching statement on the DSU's goals, it does not mean anything that Article 3 does not specifically indicate that the Panel and AB Reports should have precedential effect. In any case, an equally persuasive argument can be made that Article 3 did not specify that the Reports should *not* have precedential effect.

Secondly, the language of Articles 3(3) and 3(4) supports the contrary argument; namely, that the DSU's reports should contribute to the formation of a stable body of precedents for Members' reference and edification. In particular, Article 3(3) regards "the [*prompt*] settlement of [disputes]…" as "essential to the functioning of the WTO…" The "prompt" settlement of disputes is certainly furthered by the application of a doctrine of SD, which enables judicial economy by promoting consistency and stability in the interpretation of Members' rights before Panels and the AB.[52]

Following a related train of thought, in the *US – Stainless Steel* case mentioned earlier, the AB stated that "ensuring 'security and predictability' in the dispute settlement system, as contemplated in Article 3(2) of the DSU, implies that, absent cogent reasons, an adjudicatory body will resolve the same legal question in the same way in a subsequent case".[53] Indeed, it would be useful for "security and

[51] Marrakesh Agreement Establishing the World Trade Organization (adopted 15 April 1994, entered into force 1 January 1995) 1867 UNTS 154, Preamble.

[52] Simon Lester, 'International Decisions: WTO-Anti-dumping Agreement – "zeroing" – role of precedent – standard of review' (2008) 102(4) American Journal of International Law 834, 839.

[53] *US – Stainless Steel* (n 32) [160].

predictability" if the doctrine of SD is applicable,[54] giving voice to Members' legitimate expectations that like cases should be treated alike.

It must be acknowledged, however, that although these are good reasons for the application of SD, they do not yet lead to the *necessary* conclusion that the doctrine of SD must apply in the DSU. As such, further guidance must be sought elsewhere.

B. Doctrine of SD is implied in the DSU

The interpretation of the WTO Agreement and the DSU are both guided by "customary rules of interpretation of public international law",[55] which has been codified in the form of the Vienna Convention on the Law of Treaties (VCLT).[56] In particular, Articles 31, 32, and 33 of the Vienna Convention on the Law of Treaties.

The VCLT in its "General Rule of Interpretation" does not expressly mention the doctrine of SD nor does it describe anything similar to the doctrine. However, a few VCLT provisions appear to tacitly permit the doctrine of SD in the WTO dispute settlement context.

(I) Article 31(3)(b)

Article 31(3)(b) of the VCLT states that "any subsequent practice in the application of the treaty which establishes the agreement of the parties regarding its interpretation" should be taken into account when interpreting the treaty (i.e. the mechanism encapsulated in the DSU). In *Japan – Taxes on Alcoholic Beverages*, it was argued that adopted Reports constitute "subsequent practice", and these Reports' findings are therefore part of the "subsequent practice" of the WTO Agreement.[57] As a result, per Article 31(3)(b), Panels and ABs must take into

[54] The "cogent reasons" test introduced in *US – Stainless Steel* is one of the methods by which the doctrine of SD could be realised and applied.

[55] *US – Stainless Steel* (n 32) [39], [76], [136] and [161]. See also: WTO, 'WTO Analytical Index: Marrakesh Agreement' [279] <https://www.wto.org/english/res_e/booksp_e/analytic_index_e/wto_agree_04_e.htm#articleXVI> accessed 26 December 2017.

[56] Vienna Convention on the Law of Treaties (adopted 23 May 1969, entered into force 27 January 1980) 1155 UNTS 331.

[57] Chua (n 1) 182.

account these past findings in interpreting the WTO Agreement. This essentially transplants the doctrine of SD into the Panel and AB's decision-making process.

However, this argument was rejected in the *Japan – Alcoholic Beverages*[58] case on the grounds that "the essence of subsequent practice in interpreting a treaty has been recognised as a 'concordant, common and consistent' sequence of acts or pronouncements which is sufficient to establish a discernible pattern implying the *agreement* of the parties regarding its interpretation".[59] As such, it was determined that a single precedent would be insufficient to establish "subsequent practice"[60] such that parties' agreement to the interpretation could be inferred.

Although this appears to be a rejection of the doctrine of SD, the AB did not unequivocally decline to establish its own power to set precedents. It merely stated that isolated incidents would be insufficient "practice" to convincingly establish "agreement" as to the interpretation. Hence, where there is a sufficient sequence of cases agreeing on the same interpretation for a given clause in the WTO Agreement, the AB would likely accept the particular interpretation as conclusive since the interpretation would be taken to enjoy the acceptance of all parties, per Article 31(3)(b). In such a situation, the doctrine of SD would operate by virtue of Article 31(3)(b) VCLT read with Article 3(2) DSU, as the string of past decisions would take on precedential effect and become normally binding.

In such a situation, the doctrine of SD applies but with a caveat. An interpretation becomes "normally binding" only when there is a sufficiently long and consistent line of prior decisions concurring with the interpretation. Exactly how many Reports would be required to constitute sufficient "subsequent practice" is open to further debate. It would seem then that a persuasive argument has been made to the effect that a limited but satisfactory doctrine of SD applies, per the relevant provisions in the VCLT and the DSU.

Unfortunately, one further problem arises: what of the situation where a string of cases is built upon each other in an illegitimate practice of *de facto* SD, reaching the same interpretation of a given clause in the WTO Agreement? The line of precedents could then be traced back to a *single* case, which would run contrary to the *spirit* of a "sequence of acts" sufficient to establish "subsequent practice". Hence, it would appear that the doctrine of SD would be further circumscribed in its application to situations where each of the Panels or ABs issuing the Reports cumulatively constituting "subsequent practice" independently reach the interpretation of the relevant clause of the WTO Agreement. At the very least, the

[58] *Japan – Taxes on Alcoholic Beverages* (1996) WT/DS8/AB/R.
[59] ibid 13.
[60] Chua (n 1) 183.

Panels and ABs should not have cited prior cases as dispositive reasons in reaching their interpretations of the WTO clauses in question.

Although this argument based on Article 31(3)(b) of the VCLT leads to the conclusion that the doctrine of SD is applicable to the dispute resolution process, it leaves us with a diminished version of the doctrine of SD. Hence, the next logical step would be to ascertain if there are any other provisions capable of incorporating the full doctrine of SD into the Panel's and AB's dispute resolution process.

(II) Article 31(3)(a)

Under the same Article, the VCLT also states that "any subsequent agreement between the parties regarding the interpretation of the treaty or the application of its provisions" should be taken into account during interpretation. Similar to Article 31(3)(b), this sub-paragraph also focuses on the "agreement" of the parties, but does not require that this "agreement" be proved via the *frequency* of prior Reports reaching the same interpretation of a given clause in the treaty.

In this section, "agreement" will not be assessed with regard to agreement as to any particular interpretation of any particular clause in the WTO Agreement, as was the case when analysing the implications of Article 31(3)(b). Instead, the central question which needs to be answered is whether Members have agreed, generally, to a doctrine of SD in the WTO Agreement.

It is not contested that in Article 4(2) of the WTO Agreement, signatory states unanimously agreed to the creation of the DSB and its compulsory jurisdiction. By extension, the Members also consented to the DSU annexed to the Marrakesh Agreement, from which the DSB derives its functions and powers. Under Articles 3(1) and 17(14) of the DSU, Members agree that where they are parties to a case, they will accept the Panel's and AB's interpretations and abide by the DSB's final ruling. The problem with this is that this agreement and acceptance of the interpretation is limited to the parties in a case. Hence, the precedential value of such a case would be limited to future cases where the parties again appear as litigants.

However, it is possible to discern "subsequent agreement", not merely on the interpretation of a specific provision in a specific case, but a more general, lasting agreement as to the precedential value of Reports in general. Agreement need not be demonstrated in the form of a formal contract, and may also take the

form of unequivocal acts, for the substance is more important than the form of agreement.[61]

Recall that neither the DSU nor the WTO Agreement explicitly forbids the doctrine of SD. It is only that in Article 3(2) of the DSU, the DSB pledges not to vary the rights and obligations of Members. It has never been clear whether the DSB's interpretation of the WTO Agreement constitutes variation of these rights and obligations, or—as argued earlier—whether these are merely interpretations, clarifying the boundaries of existing rights and obligations. If it is the latter, the doctrine of SD is capable of applying, and *should* apply in order to prevent wastage of resources in litigating over the same question.

Hence, any Member should be taken to have acquiesced to the operation of a doctrine of SD if, in their own submissions to the Panel or AB, they cite past decisions in order to convince the Panel or AB to adopt the same reasoning and interpretations once again. In so doing, they have decisively waived the option to challenge that according to the DSU, past DSB rulings have no precedential effect. On the contrary, each party's great hope is that the current Panel or AB will recall their past decision and be so bound. This can thus be taken as agreement between parties that neither Article 3(2) DSU nor Article IX(2) WTO, or any other provision, prevents past Reports from taking on future precedential effect.

Nearly every member who has at some point—in any dispute, or as a third party—directed the Panel's or AB's attention to a past case's interpretation can be taken to have agreed that the treaty can be interpreted with reference to past cases. In any case, in almost every dispute, there would be consent to the doctrine of SD; thus far, there has been no party that would willingly neglect to raise past cases as authority supporting their reasoning. Every Report published by the DSB comes attached with cases cited by both parties. It would not be an exaggeration to say that most Members have acknowledged the precedential effect and value of DSB Reports by petitioning adjudicators to abide by past decisions.

Such tacit agreement that the doctrine of SD applies is further reinforced by representatives' statements outside the dispute resolution process. For example, in *Canada – Administration of the Foreign Investment Review Act*,[62] the Korean representative stated that panel reports were not limited to the specific fact scenario but "constituted a precedent"; India and other developing nations were quick to observe that the Report could only contribute to future cases where both parties were developed

[61] *United States – Measures Affecting the Production and Sale of Clove Cigarettes* WT/DS406/AB/R [267].

[62] *Canada – Administration of the Foreign Investment Review Act* (1984) GATT BISD (30th Supp) 140.

parties, and could not affect future claims from developing nations.[63] In addition, it is common for representatives of winner states to refer to successful suits as setting "precedents".[64]

Consequently, over time, most Members will have in substance implicitly agreed to apply the doctrine of SD. Where a dispute is between Members who have agreed that the doctrine of SD should apply, per Article 31(3)(a), this should be an important factor in the Panel's or AB's preliminary analysis of whether under the DSU, the doctrine of SD applies. Such subsequent agreement would weigh heavily in favour of a finding that the doctrine applies.

V. Conclusion

The doctrine of SD as it now stands balances in a continuum, between unbreakable binding precedent and mere persuasive authority. While the term *stare decisis* is admittedly a common law concept, it has its counterparts in civil law. Article 5 of the French Civil Law Code famously repudiates the concept of precedence, stating that "a judge… [is not] to dispose of the case by reference… to prior decisions". Yet, *la jurisprudence* is the result of the accumulation of a body of judgments, and in practice, 90% of French judges follow the position of the *Cour de Cassation*, creating a line of similar decisions.[65] The universality of the doctrine of SD is such that it is inevitable that the DSB will ultimately adopt a *de facto* doctrine of SD.

One of the greatest fears regarding the doctrine of SD is that it will bind future adjudicators to abide by the mistakes made by the predecessors. However, the doctrine of SD has developed over the years to take a more flexible stance on horizontal SD, which allows the apex adjudicator to deviate from its past rulings, which are only "normally binding". In any case, Article IX(2) can be seen as empowering the General Council and Ministerial Council of the WTO to veto interpretations by the AB, serving as a check on judicial power, echoing the separation of powers in national systems.

Having put these pressing worries about the doctrine at ease, would it not be better adapting to reality, acknowledging the long practice of SD and accepting the approving clues scattered in the relevant parts of the WTO Agreement, and so give the doctrine of SD formal legitimacy instead of requiring that adjudicators

[63] Refer to the Minutes of the GATT Council: GATT Council 'Council – Minutes of Meeting – Held in the Centre William Rappard on 11 October 1989' (11 October 1989) C/M/236 <https://docs.wto.org/gattdocs/q/GG/C/M236.PDF> accessed 26 December 2017.

[64] Bhala (n 15) 871.

[65] Christian Dadomo and Susan Farran, *The French Legal System* (2nd edn, Thomson Professional Publishing Canada 1996) 42.

dance around the issue? As such, this article attempts to present possible routes by which the doctrine of SD can be formally and openly inducted into the operation of the DSB. It is only when the official position of the DSB matches its actions that it can more efficiently fulfil its central function—to justly and transparently bring resolution to multinational trade disputes.

GATT Article XXI: Trade Sanctions and the Need to Clarify the Security Exceptions

Henry Federer*

I. Introduction

When a state defies international norms and threatens global peace and security, trade sanctions provide an opportunity for nations to influence the acts of others by non-violent means. By using economic diplomacy to negatively affect other economies, states can seek to deter aggression and motivate change. With the continued rise of globalisation and international free trade, utilising trade sanctions can interfere with a nation's other obligations—a World Trade Organisation (WTO) Member enacting trade sanctions against another Member would be in violation of the General Agreement on Tariffs and Trade 1994 (hereinafter, "GATT 1994")[1]—meaning that fulfilling trade obligations can be detrimental to achieving national security goals.

This article will argue that trade sanctions work and that because of this the interpretation and use of Article XXI,[2] "Security Exceptions", needs to be clarified so that WTO Members can properly use it to justify GATT 1994 violating trade sanctions made to support national security. Part II of this article will argue that trade sanctions work by demonstrating how they are an effective method for

* J.D. Candidate (2019), Queen's University Faculty of Law, Kingston, Canada; B.A. (2015), Western University, London, Canada. I would like to thank the editors of the *Cambridge Law Review* for their efforts on this article. Any errors are my own.

1 General Agreement on Tariffs and Trade (adopted 15 April 1994, entered into force 1 January 1995) 1867 UNTS 3 (hereinafter, "GATT 1994").

2 ibid, Article XXI.

shifting a state's actions without the use of violence. It will do this by examining the Libyan Terrorism and WMD Sanctions and the Iranian Nuclear Sanctions. Part III of this article will examine the history of Article XXI, looking at its text and its invocations throughout history. In the pre-GATT 1994 context, it will look at invocations by Ghana, Sweden, and the United States (US); and in the post-GATT 1994 context, it will look at invocations by the US and Nicaragua. Part IV of this article will argue that Article XXI can play an important role regarding trade sanctions and as such its interpretation and use needs to be clarified. It will do this by first showing how trade sanctions violate the GATT 1994. Next, it will analyse the competing *Ultra Vires Perspective* and *Intra Vires Perspective* of Article XXI. Finally, it will show how the leading *Ultra Vires Perspective* of Article XXI is too broad to properly be utilised to justify trade sanctions and that the *Intra Vires Perspective* should be adopted.

II. Trade Sanctions Work

Trade sanctions are a form of economic diplomacy that seeks to alter a state's activities by non-violent means with the idea that economic hardship will lead them to change course. Defined by the Council on Foreign Relations as "the withdrawal of customary trade and financial relations" in either comprehensive form, such as "prohibiting commercial activity with regard to an entire country", or a more targeted form, such as "blocking transactions of and with particular businesses, groups, or individuals",[3] countries launch trade sanctions against other countries whose conduct are against their interests; these interests can be strategic, such as targeting a country's aggression towards another country, or moral, such as targeting a country's actions that are contrary to international norms. This sort of economic diplomacy has been utilised throughout modern history, such as the Embargo Act of 1807 which saw the US launch an embargo against goods from Great Britain during the Napoleonic wars,[4] and continues to be used today, such as Countering America's Adversaries Through Sanctions Act[5] which seeks to curb North Korea's nuclear ambitions, among other things. Although regularly used, a large question that is asked about sanctions is are they successful at forcing international actors to change course? This part of the article will argue that they are successful by examining the Libyan Terrorism and WMD Sanctions and the

[3] Jonathan Masters, 'What Are Economic Sanctions?' (*Council on Foreign Relations*, 7 August 2017) <https://www.cfr.org/backgrounder/what-are-economic-sanctions#chapter-title-0-9> accessed 4 September 2018.

[4] Embargo Act of 1807 (2 Stat 451).

[5] Countering America's Adversaries Through Sanctions Act of 2017 (Public Law No. 115-44).

Iranian Nuclear Sanctions. It will also examine failed sanction attempts, noting why they failed and how they are not indicative of sanctions' effectiveness.

A. Libyan terrorism and WMD sanctions

The success of the trade sanctions implemented by the US and the United Nations (UN) demonstrate that trade sanctions work. In the 1990s, important segments of Libya's economy were targeted by sanctions with the goal of altering Libya's support for terrorism and ending its chemical and nuclear weapon programs. When the Soviet Union fell and the Cold War ended, Libya lost one of its main financial allies, which, along with a jaded view of Pan-Arabism and the weakness of African economies, led to economic hardship and instability within Libya.[6] With Libya in a weakened state, the US sought to use economic diplomacy to further exacerbate Libya's woes to force Libya to end its chemical and nuclear weapons programs and its support for terrorism.[7] Continuing work done by the Bush Administration, in 1992 and 1993 the Clinton Administration successfully persuaded the UN to put forward two resolutions that resulted in comprehensive trade sanctions against Libya for its role in the destruction of Pan Am Flight 103 and UTA Flight 772.[8] Resolution 748 sought to impact Libya's economy by denying "permission of Libyan aircraft to take off from, land in or overfly their territory if it has taken off from Libyan territory"[9] and preventing the importation of aircraft parts,[10] while Resolution 883 sought to impact Libya's economy by damaging its oil industry[11] and further damage its airline industry.[12] In order for these sanctions to be lifted, Libya would have to allow investigations into the Pan Am Flight 103 and UTA Flight 772 crashes and allow the Libyan nationals who were suspected to have caused the crashes be extradited so that justice could be served.[13] As these sanctions succeeded in damaging Libya's economy, the US acted through its own agency to further impact Libya. In 1996, Congress passed, and President Clinton signed into law, the Iran and Libya Sanctions Act of 1996,[14] with the objective of driving Libya to "end all support for acts of international terrorism and efforts to

6 Martin S Indyk, 'The Iraq War Did Not Force Gadaffi's Hand' (*Brookings Institute*, 9 March 2004) <https://www.brookings.edu/opinions/the-iraq-war-did-not-force-gadaffis-hand/> accessed 4 September 2018.

7 Bruce Jentleson, 'Coercive Diplomacy: Scope and Limits in the Contemporary World' (*The Stanley Foundation*, 2006) Policy Analysis Brief 2 <http://stanleyfoundation.org/publications/pab/pab-06CoerDip.pdf> accessed 4 September 2018.

8 ibid 4–5.

9 UNSC Res 748 (31 March 1992) UN Doc S/RES/748 4(a).

10 ibid (b).

11 UNSC Res 883 (11 November 1993) UN Doc S/RES/883.

12 ibid.

13 UNSC Res 731 (21 January 1992) UN Doc S/RES/731.

14 Iran and Libya Sanctions Act of 1996 (Public Law No. 104-172).

develop or acquire weapons of mass destruction".[15] With regards to Libya, the paramount provision of this act targeted US and non-US people or institutions that:

> (A) contributed to Libya's ability to acquire chemical, biological, or nuclear weapons or destabilizing numbers and types of advanced conventional weapons or enhanced Libya's military or paramilitary capabilities;
>
> (B) contributed to Libya's ability to develop its petroleum resources; or
>
> (C) contributed to Libya's ability to maintain its aviation capabilities.[16]

This provision essentially barred Libya from being able to develop its petroleum resources, which as a petro-state was vital to its economy, or aviation industries, which heavily impacted its ability to import or export goods. These sanctions would be lifted in return for Libya ending its support of terrorism and ending its WMD programs.[17]

In 1998, six years after the initial UN sanctions were implemented, the economic effects the UN and US trade sanctions were having on Libya's economy brought Libya to the negotiating table; Libya agreed to turn over the suspects of the Pan Am bombing in return for the suspension of the UN sanctions with the caveat that the sanctions would fully end once "the terrorism case was fully settled".[18] In 2003, Libyan officials approached the US and the United Kingdom (UK), offering to end and reveal the full extent of its WMD program in return for the US lifting its sanctions;[19] after negotiations a settlement was made that brought Libya in line with the international community with regards to nuclear weapons, with Libya reaffirming its commitments to numerous international treaties including the Treaty on the Non-Proliferation of Nuclear Weapons, the Agreement on Safeguards of the International Atomic Energy Agency (IAEA) and

[15] ibid s 3(b).
[16] ibid s 5.
[17] ibid s 3(b).
[18] Jentleson (n 7) 5.
[19] Sean D Murphy, 'US/UK Negotiations with Libya Regarding Non-Proliferation' (2004) 98 AJIL 195.

the Convention on Biological Weapons.[20] The success that trade sanctions had on altering Libya's actions demonstrate that trade sanctions can succeed in bringing states to the negotiating table and alter their actions without having to resort to violent means.

B. Iranian nuclear sanctions

Like the Libyan Sanctions, the Iranian Nuclear Sanctions are an excellent case study on how trade sanctions can successfully alter a state's actions through non-violent means. For over two decades the international community led by the US had been using trade sanctions to damage Iran's economy with the goal of driving Iran to end its nuclear weapons program. In response to the "unusual and extraordinary threat to the [US'] national security"[21] that Iran's nuclear ambitions posed, in 1995 President Clinton signed Executive Order 12957, which prohibited all US trade in Iran's petroleum industry,[22] and Executive Order 12959, which prohibited all US trade with Iran.[23] In 1996, the US' economic diplomacy against Iran was further expanded with the passage of the Iran and Libya Sanctions Act of 1996,[24] which targeted foreign companies that were investing in Iran's petroleum industry with the goal of preventing the investment.[25] Although sanctions initially started to work, leading to a suspension of Iran's uranium enrichment program,[26] following President Ahmadinejad's electoral victory in 2005, Iran ended this suspension causing the US to push the UN to use its powers under Chapter VII of the United Nations Charter[27] to enact sanctions against Iran. These resolutions

[20] Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (adopted 10 April 1972, entered into force 26 March 1975) 1015 UNTS 163 (Biological Weapons Convention); Treaty on the Non-Proliferation of Nuclear Weapons (adopted 1 July 1968, entered into force 5 March 1970) 729 UNTS 161; UNSC 'Letter dated 19 December 2003 from the Permanent Representative of the Libyan Arab Jamahiriya to the United Nations addressed to the President of the Security Council' (2003) UN Doc S/2003/1196.

[21] The President of the United States of America, 'Executive Order 12959—Prohibiting Certain Transactions with Respect to Iran' (9 May 1995) 60(89) Federal Register 24757.

[22] The President of the United States of America, 'Executive Order 12957— Prohibiting Certain Transactions With Respect to the Development of Iranian Petroleum Resources' (15 March 1995) 60(52) Federal Register 14615.

[23] UNSC Letter (n 20) s 1.

[24] Now known as the Iran Sanctions Act. See Iran Sanctions Extension Act of 2016 (Public Law No. 114-277).

[25] Iran and Libya Sanctions Act of 1996 (n 14) s 5(a).

[26] Sten Rynning, 'Europe's Emergent but Weak Strategic Culture' in Kjell Engelbrekt and Jan Hallenberg (eds), *European Union and Strategy: An Emerging Actor* (Routledge 2010) 94.

[27] Charter of the United Nations (adopted 26 June 1945, entered into force 24 October 1945) 1 UNTS XVI, Chapter VII (hereinafter, "United Nations Charter").

first started in 2006 with Resolution 1696, which demanded that Iran suspend its nuclear program[28] and Resolution 1737, which officially imposed trade sanctions,[29] and would continue to be implemented until 2014,[30] dealing considerable blows to the Iranian economy.[31] Throughout these resolutions American sanctions continued, most notably in Comprehensive Iran Sanctions, Accountability, and Divestment Act of 2010[32] and Executive Order 13590,[33] both of which targeted US and non-US persons and companies investing in or purchasing from Iran's petroleum industry.[34] Furthermore, the European Union (EU) also put forward sanctions during this period, targeting "Iranian crude oil imports to the EU, in the financial sector, including against the Central Bank of Iran, in the transport sector as well as further export restrictions, notably on gold and on sensitive dual-use goods and technology."[35]

With the effects of sanctions from the US, EU, and UN devastating its economy, Iran came forward to the negotiating table seeking an end to the sanctions. In 2013 the newly elected Iranian President, Hassan Rouhani, worked with the five permanent members of the UN Security Council and Germany, known as the P5+1, to create the agreement known as the Joint Plan of Action,[36] which saw the freezing of elements of Iran's nuclear program in return for a decrease in sanctions as both sides worked towards a long-term agreement. Following this, negotiators were able to reach a final agreement known as the Joint Comprehensive Plan of Action in July 2015.[37] Through the final agreement, Iran agreed to hand over large swathes of its nuclear weapons program and take steps that would prevent it from reinstating it for over a decade; in return for this, Iran received a loosening

[28] UNSC Res 1696 (31 July 2006) UN Doc S/RES/1696.

[29] UNSC Res 1737 (23 December 2006) UN Doc S/RES/1737.

[30] UNSC Res 2159 (9 June 2014) UN Doc S/RES/2159 (2014).

[31] Central Bank of the Islamic Republic of Iran, 'Economic Trends No 62' (2010/2011) Third Quarter 1389, 16.

[32] Comprehensive Iran Sanctions, Accountability, and Divestment Act of 2010 (Public Law No 111-195) (hereinafter, "Comprehensive Iran Sanctions").

[33] The President of the United States of America, 'Executive Order 72609—Authorizing the Imposition of Certain Sanctions With Respect to the Provision of Goods, Services, Technology, or Support for Iran's Energy and Petrochemical Sectors' (23 November 2011) 76(226) Federal Register 72609.

[34] ibid s 1; Comprehensive Iran Sanctions (n 32) s 103.

[35] Council of The European Union, 'Council Conclusions on Iran' (23 January 2012) 3142th Foreign Affairs Council Meeting 2 <http://www.europarl.europa.eu/meetdocs/2009_2014/documents/d-ir/dv/council_cnclsions_iran_/council_cnclsions_iran_en.pdf> accessed 14 September 2018.

[36] Geneva Interim Agreement between Iran and P5+1 (adopted 24 November 2013) (Joint Plan of Action).

[37] Joint Comprehensive Plan of Action between Iran and P5+1 (adopted 14 July 2015).

of sanctions from the US and the EU,[38] as well as the end of UN sanctions.[39] The success trade sanctions had on altering Iran's actions demonstrate that trade sanctions can succeed in bringing states to the negotiating table and alter their actions without having to resort to violent means.

C. Failed sanctions

Although the Libyan and Iranian sanctions demonstrate that trade sanctions can succeed in altering a state's actions, it is important to note that trade sanctions are not perfect as demonstrated by failed trade sanctions targeting Cuba and North Korea. The failures, however, can be explained through flaws in the implementation of sanctions, rather than flaws in the sanctions themselves.

The US trade sanctions towards Cuba is the leading example detractors of trade sanctions bring up when arguing for their ineffectiveness. Following the Cuban revolution which saw Cuba turn into a communist state under Fidel Castro in 1959, the US' relationship with Cuba deteriorated. With the main goal of pushing Cuba away from communism into what would amount to regime change and the secondary goals of punishing Cuba for its human rights abuses and intimidating it in response to the national security threat it posed,[40] the US has put forward numerous sanctions against the country amounting to a full "economic, commercial and financial embargo"[41] that have amounted to an estimated loss of over $100 billion over the years.[42] Although these sanctions have caused vast amounts of economic damage to Cuba's economy, they have been unsuccessful in altering Cuba's actions, which has drawn criticism towards the effectiveness of trade sanctions.

The US and international community's trade sanctions towards North Korea is another leading example which detractors of trade sanctions bring up when arguing for their ineffectiveness. In 2006, North Korea launched its first nuclear weapons test. In response, the UN passed Resolution 1695,[43] the first of many UN

[38] Ellie Geranmayeh, 'Explainer: The Iran Nuclear Deal' (*European Council on Foreign Relations*, 17 July 2015) <http://www.ecfr.eu/article/iran_explainer3070> accessed 4 September 2018.

[39] UNSC Res 2231 (20 July 2015) UN Doc S/RES/2231.

[40] See The Cuban Liberty and Democratic Solidarity (Libertad) Act of 1996 (Helms-Burton Act) (Public Law No. 104-114). See also Cuban Democracy Act of 1992 (Public Law No. 102-484).

[41] UNGA Res 72/4 (10 November 2017) UN Doc A/RES/72/4.

[42] Daniel Trotta, 'Cuba Estimates Total Damage of U.S. Embargo at $116.8 Billion' *Reuters* (Toronto, 9 September 2014) <https://www.reuters.com/article/us-cuba-usa/cuba-estimates-total-damage-of-u-s-embargo-at-116-8-billion-idUSKBN0H422Y20140909> accessed 4 September 2018. See also Hildy Teegen, Hossein Askari, John Forrer, and Jiawen Yang, 'Economic and Strategic Impacts of U.S. Economic Sanctions on Cuba' (2002) Working Paper 13 <https://www2.gwu.edu/~clai/working_papers/Teegen_Hildy_02-03.pdf> accessed 4 September 2018.

[43] UNSC Res 1965 (15 July 2006) UN Doc S/RES/1695.

trade sanctions against the country. Although Resolution 1695 was more targeted towards the nuclear program, recent resolutions, such as Resolution 2375, put forward more comprehensive sanctions that target the country's whole economy.[44] During this time the US has also implemented sanctions against North Korea, including the North Korea Sanctions and Policy Enhancement Act of 2016, which targeted its export industries.[45] Even with this economic damage these sanctions have caused North Korea, including an estimate of $500 million a year,[46] North Korea has not stopped its nuclear program, once again drawing criticism towards the effectiveness of trade sanctions.

Although the trade sanctions implemented against Cuba and North Korea have not been successful, there are reasons that explain this ineffectiveness better than the conclusion that trade sanctions are ineffective. First and foremost, sanctions that put forward unrealistic requests will be unsuccessful. In the Cuban example, the US was effectively demanding regime change by demanding for the end of the communist, while in the North Korea case, the world is effectively asking it not to pursue the only tools that can assure its survival by asking the country to stop working towards nuclear weapons; as neither country would agree to these outcomes, the sanctions have been ineffective. With both Libya and Iran, the outcomes sought did not have such far reaching consequences for the decision makers in each respective country—obtaining nuclear weapons were not viewed essential to the respective regimes—making it a far more sensible request.

Secondly, as in Cuba's case, sanctions that are based on unreasonable premises will be ineffective. Whereas sanctions against Libya and Iran were based in national security and a desire to promote non-proliferation, sanctions against Cuba were based on an anti-communist ideology rather than a more substantive goal. When a sanctioned country cannot see the rationale behind trade sanctions, they are less likely to agree.

Lastly, in the North Korea case, sanctions take time to work—as noted by diplomat and Professor Victor Cha, "sanctions don't work until they do".[47] Looking to both Libya and Iran, it is important to note that it took years for these countries to come to the negotiating table following the imposition of sanctions. Although this may seem like a long time, alternatives to trade sanctions also require time. For example, using a means like war to alter a state's actions can last over a decade

[44] UNSC Res 2375 (11 September 2017) UN Doc S/RES/2375.

[45] North Korea Sanctions and Policy Enhancement Act of 2016 (Public Law No. 114-122).

[46] United States Mission to the United Nations, 'Fact Sheet: Resolution 2375 (2017) Strengthening Sanctions on North Korea' (11 September 2017) <https://usun.state.gov/remarks/7969> accessed 14 September 2018.

[47] Colin Quinn, 'The CSIS Podcast: Defusing North Korea' (7 July 2017) Center for Security Studies <https://www.csis.org/podcasts/csis-podcast/defusing-north-korea-1>.

such as is the case with the Iraq War. Even though UN sanctions have been in effect against North Korea since 2006, only recently have more comprehensive sanctions that have sought to impact all North Korea's economy come into effect, meaning that we must continue to wait before we can fully judge their effectiveness. The intricacies of the unsuccessful Cuban and North Korean sanctions described above demonstrate that trade sanctions are not ineffective, but how they are applied can have a large impact on their success.

III. History of Article XXI

Article XXI of the GATT 1994, titled "Security Exceptions", allows countries to violate their WTO obligations in the pursuit of national security. The text of the article states:

> Nothing in this Agreement shall be construed
>
> a. to require any contracting party to furnish any information the disclosure of which it considers contrary to its essential security interests; or
>
> b. to prevent any contracting party from taking any action which it considers necessary for the protection of its essential security interests
>
> (i) relating to fissionable materials or the materials from which they are derived;
>
> (ii) relating to the traffic in arms, ammunition and implements of war and to such traffic in other goods and materials as is carried on directly or indirectly for the purpose of supplying a military establishment;
>
> (iii) taken in time of war or other emergency in international relations; or
>
> c. to prevent any contracting party from taking any action in pursuance of its obligations under the United Nations Charter for the maintenance of international peace and security.[48]

With its origin in the GATT 1994's original precursor, the General Agreement on Tariffs and Trade 1947,[49] the security exception was included to balance

[48] GATT 1994 (n 1) Article XXI.

[49] General Agreement on Tariffs and Trade (adopted 30 October 1947, entered into force 1 January 1948) 55 UNTS 187 (hereinafter, "GATT 1947").

"national security and national sovereignty on the one hand, and the need to promote commerce and to protect an open trading system on the other".[50]

The most important part of the security exceptions is Article XXI:(b)(iii) because of its broadness and vagueness and the controversy these attributes have attracted. It states that "nothing in this agreement shall be construed… to prevent any contracting party from taking any action which it considers necessary for the protection of its essential security interests... taken in time of war or other emergency in international relations".[51] As Peter Lindsay notes, these attributes are rooted in the fact that the GATT 1994 has not defined terms such as "'considers necessary,' 'essential security interests,' 'time of war,' and 'emergency in international relations'",[52] creating ambiguity on how it should be invoked and applied. The controversy this fact has attracted include the view that the exception can be called to justify any measure a Member views as supporting national security, bestowing WTO Members the power to invalidate any WTO obligation and creating a hesitancy to challenge any country's use of the article as it would amount to challenging a country's ability to decide their own national security interests. Looking to the use of the Article XXI in pre- and post-GATT 1994 contexts demonstrates the variety of circumstances where Article XXI has been applied.

A. Pre-GATT 1994

Before the GATT 1994 there were only a handful of occasions where GATT 1947 signatories invoked Article XXI to justify their violation of their treaty obligations. One early instance was Ghana's invocation of it in 1961. Here, Ghana invoked the Article to justify its boycott of Portuguese goods; it was boycotting the goods under the belief that it would put pressure on Portugal and lessen the danger of the Angolan Independence War, which Ghana argued posed a danger to "the peace of the African continent".[53] When invoking this Article Ghana noted that, in its interpretation of it, "each contracting party was the sole judge of what was necessary in its essential security interest", and that as such there could "be

[50] Alan S Alexandroff and Rajeev Sharma, 'The National Security Provision—GATT Article XXI' in Patrick FJ. Macrory, Arthur E Appleton, and Michael G Plummer (eds), *The World Trade Organization: Legal, Economic and Political Analysis* (Springer 2005) 1572.

[51] GATT 1994 (n 1) Article XXI:(b)(iii).

[52] Peter Lindsey, 'The Ambiguity of GATT Article XXI: Subtle Success or Rampant Failure?' (2003) 52 DLJ 1277, 1278.

[53] GATT Contracting Parties, 'Summary Record of The Twelfth Session' (21 December 1961) SR 19/12, 196.

no objection to Ghana regarding the boycott of goods as justified by security interests",[54] demonstrating the view that one can utilise Article XXI unilaterally.

Another instance in the history of Article XXI in the pre-GATT 1994 context was Sweden's invocation of it in 1975 when they instituted a global import quota system for certain footwear under tariff headings ex 64.01 and ex 64.02.[55] Here Sweden argued that the quota was necessary because:

> [D]ecrease in domestic [footwear] production has become a critical threat to the emergency planning of Sweden's economic defence as an integral part of the country's security policy. This policy necessitates the maintenance of a minimum domestic production capacity in vital industries. Such a capacity is indispensable in order to secure the provision of essential products necessary to meet basic needs in case of war or other emergency in international relations.[56]

Many countries expressed concern with Sweden's decision, arguing that Sweden was not making this decision to further national security, but rather—noting the lack of economic justification in regards to national security—that its decision was to help deter the effects of the present global recession on its economy.[57] Although the invocation was never challenged and the decision was later revoked in 1977,[58] this provides an excellent example of an invocation of Article XXI to justify suspect national security measures.

One last important invocation of Article XXI in the pre-GATT 1994 context was when the US invoked it in 1985. It declared that it was prohibiting all imports from Nicaragua, with the purpose of undermining the Sandinista government.[59] On the day the embargo was declared, President Ronald Reagan stated that "the policies and actions of the Government of Nicaragua constitute an unusual and extraordinary threat to the national security and foreign policy of the United States and hereby declare a national emergency to deal with that threat."[60] In Council, Nicaragua argued that the US's measure violated Articles I, II, V, XI, XIII and Part IV of the GATT 1947, and that the actions were not being taken

[54] ibid.

[55] GATT Council, 'Minutes of Meeting' (10 November 1975) C/M/109, 8.

[56] ibid.

[57] ibid 9.

[58] Group of Negotiations on GATT Articles, 'Article XXI: Note by the Secretariat' (18 August 1987) MTN.GNG/NG7/W/16, 7.

[59] The President of the United States of America, 'Executive Order 12513—Prohibiting trade and certain other transactions involving Nicaragua' (7 May 1985) 50 Federal Register 18629.

[60] ibid.

in the name of national security.[61] Arguing that the actions taken were arbitrary, Nicaragua requested that a panel be set up to examine the issue, showing that countries believe Members can judge if other countries' security measures fall under Article XXI; the panel was however inconclusive.[62]

B. Post-GATT 1994

Since the GATT 1994 came into action, there have been two major instances where Article XXI has been invoked, but neither of them were reviewed by a panel. The first instance was when the US enacted the Helms-Burton Act on 12 March, 1996.[63] Launched in response to the Cuban government shooting down two civil aircrafts that were conducting a search and rescue mission,[64] the act increased sanctions against Cuba with the stated goal of bringing about a transition to a representative democracy and market economy in Cuba[65] to promote the US' national security.[66] One of its more contentious passages put forward "liability for trafficking in confiscated property claimed by [US] nationals".[67] In practice this section effectively punished any company doing business with Cuba, essentially forcing companies to choose between doing business with America or Cuba. The act was met with swift condemnation by the international community, including many of the US' allies like the EU and Canada, largely because of the effect it would have on "businesses that had recently entered, or were planning to establish, joint ventures in Cuba".[68] On 8 October, 1996, the EU requested a WTO panel to find the Helms-Burton Act inconsistent with the US' obligations under the GATT 1994,[69] and on 20 February 1997 a panel was called.[70] Initially indicating that it would put forward an Article XXI defence should the dispute go to a panel, the US argued that the WTO "lack[ed] competence to adjudicate a national security issue" and choose to reject any participation in the panel.[71] As the EU was preparing to

61 GATT Council, 'Minutes of Meeting' (28 June 1985) C/M/188, 4.
62 *United States – Trade Measures affecting Nicaragua* (1986) GATT BISD L/6053.
63 Helms-Burton Act (n 40).
64 UNSC 'Note by the Secretary General' (1 July 1996) UN Doc S/1996/509.
65 Helms-Burton Act (n 40) s 2(3), s 3.
66 ibid s 3.
67 ibid s 301.
68 Rene E Browne, 'Revisiting National Security in an Interdependent World: The GATT Article XXI Defense after Helms-Burton' (1997) 86 GLJ 405, 407.
69 WTO, *United States – The Cuban Liberty and Democratic Solidarity Act* (1996) WT/DS38/2 (Panel Request).
70 Browne (n 68) 407.
71 ibid 408.

schedule its first submission to the panel, an agreement was reached, preventing the need to analyse Article XXI.

The second post-GATT 1994 Article XXI instance took place in 2000 when Nicaragua enacted Law 325 of 1999, which enacted sanctions against Honduras and Colombia in the form of a tariff on goods and services.[72] These sanctions were put forward as part of a long, ongoing territorial dispute between the countries; Colombia and Honduras had just ratified a treaty on maritime boundaries that would impact areas claimed by Nicaragua.[73] In response to these sanctions, Colombia alleged that they were in violation of Articles I and II of the GATT 1994 and requested that the WTO set up a panel. Nicaragua subsequently invoked Article XXI as a defence for its sanctions, stating that this Article confirms "the inherent right of a State to protect its security and constitute an exception to the multilateral trade rules," meaning "these provisions cannot be subjected to an examination by a panel."[74] Although the Dispute Settlement Body (DSB) agreed to set up a panel, the Chairman of the DSB met with both parties and found a settlement, thus avoiding the need to analyse Article XXI through a WTO dispute mechanism.[75]

C. Recent context

Although the use of Article XXI has yet to be resolved at a panel in the post-GATT 1994 era, it is continued to be invoked today as justification for trade sanctions imposed between WTO Members. Most recently this has taken place regarding Russia and Qatar sanctions. In 2014 Russia invaded the Ukrainian Autonomous Republic of Crimea and started supporting separatists in Eastern Ukraine, moves that struck fear in many countries; drawing comparisons to Hitler's aggressiveness in the late 1930s to Eastern European countries including Czechoslovakia and Poland, many feared that the Crimea invasion would be a precursor to Russia attempting reassert control over former Soviet states.[76] These moves launched near universal condemnation by the West and the start of various

[72] WTO, *Nicaragua – Measures Affecting Imports from Honduras And Colombia* (2000) WT/DS188/2/Corr.1 (Panel Request).

[73] *Territorial and Maritime Dispute between Nicaragua and Honduras in the Caribbean Sea (Nicaragua v Honduras)* (Judgment) [2007] ICJ Rep 659, 683.

[74] WTO, *Nicaragua – Measures Affecting Imports from Honduras And Colombia* (Statements by Nicaragua) (2000) WT/DSB/COM/5/Rev.1, 2.

[75] Eric J Lobsinger, 'Diminishing Borders in Trade and Terrorism: An Examination of Regional Applicability of GATT Article XXI National Security Trade Sanctions' (2006) 13 ILSA JI&CL 99, 107.

[76] Steven Chase, 'Harper Compares Russia's Crimea Moves to Third Reich aggression' *Globe and Mail* (Toronto, 4 March 2014) <https://www.theglobeandmail.com/news/politics/canada-suspends-military-activities-with-russia/article17289679/> accessed 4 September 2018.

economic sanctions against Russia by parties such as the US, Canada, and the EU.[77] These sanctions included the banning of arms and arms related sales to Russia along with restrictions on financial services,[78] and have been viewed as *prima facie* violations of Article I and Article III of the GATT 1994[79] that could be defended by invoking Article XXI because they were enacted to deter Russian aggression. Although Russia has yet to request a panel to analyse these sanctions, they, along with Russia's responsive trade sanctions against the West, have brought Article XXI back to prominence.

In June 2017, the United Arab Emirates (UAE), Saudi Arabia, and Bahrain imposed trade sanctions against Qatar, contending that the move is to target Qatar's support of terrorism and goals of destabilising the region.[80] Qatar has since asserted that these sanctions are in violation of Articles I, V, X:1, X:2, XI, and XIII of the GATT 1994[81] and, following the UAE's decision not to participate in consultations with Qatar, has led Qatar to request a panel; a panel was set up on 22 November, 2017, to review the dispute.[82] Throughout this time the sanctioning countries have argued that their actions fall under Article XXI and that as such the matter did not fall under the competence of the WTO,[83] reviving questions on whether the WTO has jurisdiction to review Article XXI.

IV. Article XXI and Trade Sanctions

Article XXI can play a useful role regarding trade sanctions, but to properly play this role it needs to be clarified. As demonstrated by the Libyan Terrorism

[77] See Council Decision 2014/145/CFSP of 17 March 2014 concerning restrictive measures in respect of actions undermining or threatening the territorial integrity, sovereignty and independence of Ukraine [2014] OJ L 78. See also: The President of the ited States of America, 'Executive Order 13660—Blocking Property of Certain Persons Contributing to the Situation in Ukraine' (10 March 2014) 79(46) Federal Register 13493.

[78] Rishika Lekhadia, 'Can the West Justify its Sanctions against Russia under the World Trade Law?' (2015) IJIEL 151, 157.

[79] ibid 159.

[80] Tamara Qiblawi, Mohammed Tawfeeq, Elizabeth Roberts and Hamdi Alkhshali, 'Qatar Rift: Saudi, UAE, Bahrain, Egypt Cut Diplomatic Ties' *CNN* (New York, 27 July 2017) <http://www.cnn.com/2017/06/05/middleeast/saudi-bahrain-egypt-uae-qatar-terror/index.html> accessed 4 September 2018.

[81] WTO, *United Arab Emirates – Measures Relating to Trade in Goods and Services, and Trade-Related Aspects of Intellectual Property Rights* (2017) WT/DS526/1.

[82] WTO, *United Arab Emirates – Measures Relating to Trade in Goods and Services, and Trade-Related Aspects of Intellectual Property Rights* (2017), WTO Doc WT/DS526/2 (Request for Panel) (hereinafter, "*UAE – Trade*").

[83] WTO Council for Trade in Goods, 'National Security Cited in Two Trade Concerns at Goods Council Meeting' (*WTO*, 30 June 2017) <https://www.wto.org/english/news_e/news17_e/good_10jul17_e.htm> accessed 4 September 2018.

and WMD Sanctions and the Iranian Nuclear Sanctions, trade sanctions can successfully alter a state's actions by non-violent means. When sanctions are enacted for national security purposes, Article XXI can play a vital role in supporting the violations of the GATT 1994 these sanctions cause. However, as demonstrated by the historical analysis of Article XXI in both the pre- and post-GATT 1994 contexts, the invocation of Article XXI has been utilised for both political and security reasons—the latter of which is okay, while the former is not—and in an opaque manner. As the world continues to move towards non-violent methods to solve strategic differences,[84] and globalisation and free trade continues to spread,[85] Article XXI can and should play a vital role in our world.

A. The GATT 1994 and trade sanctions

When trade sanctions are put against WTO Members, the GATT 1994 is violated. Although some sanctions can break multiple Articles, they all break Article I, the "General Most-Favoured-Nation Treatment" provision.[86] Article I:1 states:

> With respect to customs duties and charges of any kind imposed on or in connection with importation or exportation or imposed on the international transfer of payments for imports or exports, and with respect to the method of levying such duties and charges, and with respect to all rules and formalities in connection with importation and exportation, and with respect to all matters referred to in paragraphs 2 and 4 of Article III,* any advantage, favour, privilege or immunity granted by any contracting party to any product originating in or destined for any other country shall be accorded immediately and unconditionally to the like product originating in or destined for the territories of all other contracting parties.[87]

As noted by the Appellate Body in *Canada – Certain Measures Affecting the Automotive Industry* (hereinafter, "*Canada – Autos*"),[88] when a country "has granted

[84] See Steven Pinker, *The Better Angels of Our Nature* (Viking 2011).

[85] See WTO, 'World Trade Statistical Review 2017' (WTO, 12 October 2017) <https://www.wto.org/english/res_e/statis_e/wts2017_e/wts17_toc_e.htm> accessed 4 September 2018. But see also Luke Kawa, 'Global Trade Growth is About to Roll Over' *Bloomberg* (New York, 3 May 2017) <https://www.bloomberg.com/news/articles/2017-05-03/global-trade-growth-is-about-to-slow-morgan-stanley-says> accessed 4 September 2018.

[86] GATT 1994 (n 1), Article I.

[87] ibid.

[88] WTO, *Canada – Certain Measures Affecting the Automotive Industry* (31 May 2000) WT/DS139/AB/R-WT/DS142/AB/R.

an 'advantage' to some products from some Members that [a Member] has not 'accorded immediately and unconditionally' to 'like' products 'originating in or destined for the territories of all other Members,'"[89] Article I:1 is violated. In practice, this means that countries cannot discriminate against other countries. As the nature of trade sanctions discriminate against one country, Article I is violated when they are enacted.

Looking to pre-GATT 1994 examples, sanctions like Ghana's boycott of Portuguese goods in 1961 and the US' boycott of Nicaraguan goods in 1985 represent clear discrimination. Here, the trade measures put forward by Ghana and the US discriminated against Portugal and Nicaragua respectively by according other WTO Members not subject to the sanctions more favourable trade treatment, meaning Article I is violated. Looking to the post-GATT 1994 example of the sanction enacted by Nicaragua against Colombia and Honduras in 1999, a similar conclusion can be made; Nicaragua's sanctions afforded other WTO Members more favourable trade treatment, thus breaching Article I.

Looking to the post-GATT 1994 example of the sanctions put forward by the US in the Helms-Burton Act in 1996, the breach is less clear. Through the Helms-Burton Act the United States was not sanctioning specific companies based on their origin, which would constitute a clear violation of the GATT 1994, but were sanctioning companies based on their actions. As the sanction is origin-neutral in law, the US could have tried to argue that it was not a violation. This distinction between *de jure* and *de facto* discrimination was however analysed in *Canada – Autos*, where it was ruled that Article 1 covers *de facto* discrimination.[90] As the act indirectly targeted nations that did business with Cuba, *de facto* discrimination would be present, and Article I would be violated.

Although Libya and Iran are not members of the WTO, if they were, the actions taken against them would have also violated the GATT 1994 because of their discriminatory nature. When looking at these sanctions it is important to note that the United Nations sanctions could have been justified under Article XXI:(c)[91] since they were put forward under the Chapter VII of the United Nations

[89] ibid 81.
[90] ibid 78.
[91] GATT 1994 (n 1) Article XXI(c).

Charter.[92] The non-UN sanctions, however, which played a vital role in bringing about change, would not have been.

As these numerous examples show, trade sanctions violate the GATT 1994. When faced with national security threats that do not amount to a breach "international peace and security",[93] such as the regional threats of Russia against Ukraine and Qatar against other gulf states, countries who choose to use economic diplomacy to deal with misbehaving WTO Members are faced with few options; they can either hope that their violations of the GATT 1994 can be justified by other means or open a proverbial can of worms and attempt to justify their acts under Article XXI:(b)(iii). Furthermore, as China and Russia, two states who have recently worked against international norms,[94] have veto power against any invocation of Chapter VII,[95] UN sanctions will not always be available thus removing the ability to use Article XXI:(c). As such, Article XXI needs to be clarified so countries can properly utilise trade sanctions to promote and reach national security objectives.

B. Competing Views of Article XXI

Article XXI:(b)(iii) needs to be clarified so that it can be properly invoked when WTO Members need to justify GATT 1994 violating trade sanctions they have enacted in the name of national security. As noted in the Decision Concerning Article XXI of The General Agreement, the contracting parties have not made a formal interpretation.[96] Looking to the history of Article XXI as analysed in Part III of this article, there are two main perspectives of Article XXI. The first, more popular perspective is the '*Ultra Vires Perspective*'. This perspective puts forward that the WTO does not have the jurisdiction to review the use of Article XXI. Also known as the self-judging perspective, the *Ultra Vires Perspective* states that each WTO Member is the sole judge of what actions are in their national security

[92] The UN Sanctions would have been justified because Chapter VII states that 'the Security Council shall determine the existence of any threat to the peace, breach of the peace, or act of aggression' and can take action that 'may include complete or partial interruption of economic relations' in order to get them comply with the United Nations requests, and Article XXI:(c) states the GATT 1994 cannot "prevent any contracting party from taking any action in pursuance of its obligations under the United Nations Charter for the maintenance of international peace and security".

[93] United Nations Charter (n 27) Chapter 29 Article 39.

[94] This is demonstrated by Russia's annexation of Crimea and China's island building in the South China Sea.

[95] This is because they are members of the United Nations Security Council.

[96] GATT, 'Decision Concerning Article XXI of The General Agreement' (1982) L/5426.

interest, essentially meaning that Article XXI can be used as a trump card to justify any GATT 1994 violation.

The *Ultra Vires Perspective* is first supported by interpreting Article XXI "in good faith in accordance with the ordinary meaning to be given to the terms of the treaty in their context and in the light of its object and purpose" as directed by Article 31 of the Vienna Convention on the Law of Treaties.[97] Although the purpose of the treaty is to promote trade, the purpose of the Article is to promote national security, and taken together, their purpose is to balance "national security and national sovereignty… and the need to promote commerce and to protect an open trading system."[98] With this object and purpose in mind, a plain reading of the text: "it considers… essential security interests",[99] supports the *Ultra Vires Perspective*. The phrase "it considers" suggests that only the invoking Member can judge what is in their national security interest, which supports the purpose of the act of calming any fears or doubts that the GATT 1994 would impact a Member's ability to defend its nation.

This reading of the text is supported by scholar Raj Bhala who notes that "the implication of the word 'it' indicates that no WTO Member, nor group of Members, and no WTO panel or other adjudicatory body, has any right to determine whether a measure taken by a sanctioning Member satisfies the requirements".[100] Furthermore, the lack of an Article XX *chapeau*,[101] which states that "such [Article XX] measures are not applied in a manner which would constitute a means of arbitrary or unjustifiable discrimination between countries where the same conditions prevail, or a disguised restriction on international trade"[102] further supports this reading of the text because its absence means that nations invoking the Article do not need to establish the absence of unjustifiable discrimination or a disguised appearance as they do when invoking Article XX.[103] In his piece "The Self-Judging WTO Security Exception",[104] Alford put forward why Article XXI

[97] Vienna Convention on the Law of Treaties (adopted 23 May 1969, entered into force 27 January 1980) 1155 UNTS 331, Article 31.

[98] Alexandroff and Sharma (n 50) 1572.

[99] GATT 1994 (n 1), Article XXI:(b).

[100] Raj Bhala, 'National Security and International Trade Law: What the GATT Says, and What The United States Does' (1998) 19 UPJIL 263, 268–269. See also: Raj Bhala, *International Trade Law: Interdisciplinary Theory and Practice* (4th edn, Carolina Academic Press 2015) 581.

[101] GATT 1994 (n 1), Article XX.

[102] ibid.

[103] See WTO, *European Communities – Measures Affecting Asbestos and Products Containing Asbestos* (12 March 2001) WT/DS135/AB/R (EC – Asbestos).

[104] Roger P Alford, 'The Self-Judging WTO Security Exception' (2011) 3 ULR 697.

must be viewed as self-judging, noting that the lack of the *chapeau*[105] signifies an intention for the article to be self-judging.

The second reason why the *Ultra Vires Perspective* is supported is because the language used by nations invoking Article XXI. When Ghana invoked the Article in 1961, it noted that "each contracting party was the sole judge of what was necessary in its essential security interest".[106] Similarly, when the US invoked the Article in defence of the Helms-Burton Act, it noted that the WTO "lack[ed] competence to adjudicate a national security issue",[107] and more recently Bahrain noted that judging Article XXI did not fall under the competence of the WTO.[108] As three countries who have invoked the Article in different national security situations throughout the history of the WTO have come to the same conclusion on how Article XXI should be judged, these invocations suggest that WTO Members view Article XXI through the *Ultra Vires Perspective*.

Although the *Ultra Vires Perspective* is the leading view on Article XXI, there is support for the contrary '*Intra Vires Perspective*'. The *Intra Vires Perspective* puts forward the view that Article XXI is within the WTO's jurisdiction and as such the invocation of it can be reviewed by a panel. This view is split into two camps—the first camp argues that the WTO can only review if the Article was invoked in good faith, while the other argues that the WTO can review all elements of the invocation[109]—but both are part of the *Intra Vires Perspective*. This view is first supported by the language used by the drafters of Article XXI. During a session of the Committee for the United Nations Conference on Trade and Employment in 1947, specific language was chosen to ensure that the Article would not "permit anything under the sun",[110] suggesting that drafters wanted safeguards to prevent abuse, such as the ability for invocations to be reviewed.

This interpretation is also supported by the act of challenging violations of the GATT 1947 that have used Article XXI for justification. By requesting a panel to review impugned trade measures, states are essentially challenging a party's right to unilaterally justify GATT 1947 violations in the name of national security. Looking to the language used by challengers, this view is supported. When Nicaragua challenged the trade sanctions enacted against them by the US in 1985,

[105] ibid 705.
[106] Summary Record of The Twelfth Session (n 53) 196.
[107] Browne (n 68) 408.
[108] WTO Council for Trade in Goods (n 83).
[109] See Alford (n 104) 704.
[110] UN Preparatory Committee of the International Conference on Trade and Employment, 'Report of the 2nd Session of The Preparatory Committee of The United Nations Conference on Trade and Employment (24 July 1947) EPCT/A/PV/33, 20.

they noted that Article XXI could not be applied in an arbitrary fashion[111] and that since the matter at hand involved international trade the GATT 1947 "should express a view on this issue."[112]

Similarly, when the EU challenged the US' GATT 1994 violations stemming from the Helms-Burton Act, they viewed it as "an impermissible restriction on international trade",[113] demonstrating that Article XXI cannot be permissible at every invocation. Lastly, looking to the recent Qatari sanctions, in requesting a panel to investigate the trade sanctions put against them, "Qatar said [an Article XXI] defence could not be self-regulated as that would alter the balance of Members' rights under the WTO agreements."[114] These competing perspectives create difficulty when Article XXI is invoked as it leads to confusion on how they should be used. As such, Article XXI needs to be clarified.

C. The way forward

The *Ultra Vires Perspective* is flawed because having a pure self-judging view opens Article XXI up to abuse and devalues it when it is invoked for legitimate security reasons; as history has shown, the Article has been utilised in both legitimate and illegitimate contexts and as such needs to be clarified to the *Intra Vires Perspective* to ensure use for the former is protected while use for the latter is not. Looking to history, two main illegitimate uses of Article XXI come to mind. First is the US trade embargo against Nicaragua which was widely regarded as piece of economic coercion for political reasons, rather than defensive reasons. Stating that the sanctions were put forward "not a matter of national security but one of coercion",[115] Nicaragua pointed out the absurdity of the most powerful nation in the world being fearful of a small South American country and implied that the US simply trying to coerce Nicaragua away from its communist system of government. Another instance of an illegitimate use was the Helms-Burton Act. Although Cuba historically posed a national security threat to the US, notably during the Cuban Missile Crisis, when it invoked Article XXI in 1996 to justify trade sanctions this was no longer the case. Even though the US pointed to the downing of two search and rescue planes over Cuba as the reason for the act, in reality this was just used as a false pretence to enact the sanctions; the Helms-

[111] Minutes C/M/188 (n 61) 16.
[112] ibid.
[113] Browne (n 68) 407.
[114] WTO, 'Qatar Seeks WTO Panel Review of UAE Measures on Goods, Services, IP rights' (*WTO*, 23 October 2017) <https://www.wto.org/english/news_e/news17_e/dsb_23oct17_e.htm> accessed 4 September 2018.
[115] Minutes C/M/188 (n 61) 4.

Burton Act was simply a repeat of an earlier act that failed to pass congress in 1995 before the downing.[116]

Looking to history there are multiple examples of where Article XXI has been invoked or could have been invoked for legitimate national security purposes. One example is Nicaragua's invocation in 1999 in response to legitimate territorial threats put forward by Colombia and Honduras; the sanctioned nations' new treaty stated that parts of Nicaragua's maritime boundaries belonged to them, giving Nicaragua legitimate reason to utilise economic diplomacy to try to alter their actions. Although not invoked during the Libyan and Iranian sanctions as these states are not WTO Members, Article XXI could have been legitimately utilised to justify the trade sanctions the US put against them; both Libya and Iran were seeking nuclear weapons and the US had good reason to utilise economic diplomacy to prevent them from obtaining them.

As the current leading view of Article XXI allows states to justify these illegitimate sanctions, the *Intra Vires Perspective* should be adopted. This can be done several ways such as through bold jurisprudence in the upcoming Qatar panel[117] or through an amendment to Article XXI. This amendment would be instituted through the WTO amendment procedure put forward in Article X of the Marrakesh Agreement.[118] Every two years the WTO Ministerial Conference meets where any Member of the WTO may submit a proposal to amend the GATT 1994; for an amendment to be adopted a two-thirds majority of the Members are needed.[119] Yet, in practice this procedure is very difficult to complete.[120] Looking to the Doha Development Round, attempts to amend the GATT 1994 to lower tariffs have been unsuccessfully negotiated since 2001.[121]

Should an amendment successfully come to fruition, there are many options for how to change Article XXI. One idea would be adding an Article XX *chapeau* to Article XXI, which would ensure that measure neither led to arbitrary or unjustifiable discrimination, nor constituted a disguised restriction on international trade.[122] Another idea would be to add text to article XXI such as "All Article XXI

[116] Helms-Burton Act. The bill was introduced on 14 February 1995, and the Cuban plane crash occurred on 24 February 1996.

[117] *UAE – Trade* (n 82).

[118] Marrakesh Agreement Establishing the World Trade Organization (adopted 15 April 1994, entered into force 1 January 1995) 1867 UNTS 154 (hereinafter, "the Marrakesh Agreement") Article X.

[119] ibid, Article X:1.

[120] See Macrory, Appleton, and Plummer (n 50) 85.

[121] WTO Ministerial Conference, 'Ministerial Declaration' (20 November 2001) WT/MIN(01)/DEC/1 (Ministerial Declaration).

[122] See *EC – Asbestos* (n 103). See also WTO, *European Communities – Measures Prohibiting The Importation and Marketing of Seal Products* (22 May 2014) WT/DS400/AB/R-WT/DS401/AB/R.

invocations are subject to review by the WTO. Article XXI invocations, however, are presumed to be legitimate and the onus is on the challenging Member to show otherwise." This text would essentially institute the *Intra Vires Perspective*, while also creating a significant safeguard for invokers and a barrier for challengers. A last idea would be to add a new section to Article XXI. Article XXI:(d) could state "to prevent any contracting party from enacting trade sanctions taken to deter territorial and cyber threats, the support of terrorism, and the development of weapons of mass destruction." This sort of language puts forward objective standards that the WTO can look to when judging an Article XXI invocation, while giving Member States significant leeway to violate the GATT 1994 through trade sanctions.

V. Conclusion

As demonstrated, trade sanctions work and as such the interpretation and use of Article XXI, "Security Exceptions", needs to be clarified so that WTO Members can properly use it to justify GATT 1994 violating trade sanctions made to support national security. By showing how trade sanctions can effectively shift a state's action without the use of violence through the examples of Libyan Terrorism and WMD Sanctions and the Iranian Nuclear Sanctions, this article showed that trade sanctions work. Drawing on the effectiveness of trade sanctions, this article argued that Article XXI needs to be clarified so that WTO Members can adequately use it to justify GATT 1994 violations made in the name of national security. By showing the opacity of Article XXI through historical examples of its invocation as well as the flaws of the leading *Ultra Vires Perspective*, this article has demonstrated that changes need to be made to Article XXI so that the better *Intra Vires Perspective* can be adopted.

The Price of Tea in China: Analogue Price Methodology in Anti-Dumping Investigations After the Expiry of Section 15(a)(ii) of China's WTO Accession Protocol

Josephine Williams*

I. Introduction

The expiry of Section 15(a)(ii) of China's Protocol of Accession to the World Trade Organization (WTO) leaves many unsettled legal questions. Prime among these are: (1) whether or not the remaining elements of Section 15 continue in force after 11 December 2016; (2) whether importing WTO Members are still permitted to treat China as a non-market economy under their national laws after 11 December 2016; and (3) whether China continues to be a non-market economy. This article examines each of these questions and ultimately finds that authority to determine market economy status is delegated to WTO Members under Article VI of the 1994 General Agreement on Tariffs and Trade (hereinafter "GATT 1994"). WTO Members are permitted to use alternative price methodologies under national laws when market-determined, comparable prices in the ordinary course of trade are unavailable.

II. Background

Dumping occurs when a product is sold below its cost of production or below its price on the domestic market.[1] To calculate dumping margins, investigators subtract the price at which a product is sold in the export market from the

* B.A. Wellesley College, J.D. University of California, Los Angeles. Special thank you to my UCLA Law mentor, Asli Bali and Tim Billbright of Georgetown Law.

1 Peter Van den Bossche, *The Law and Policy of the World Trade Organization: Text, Cases and Materials* (2nd edn, Cambridge University Press 2008) 698–702.

product's price in the domestic market. Dumping margin calculations are, however, complicated in the case of non-market economies. For the purposes of this article, non-market economies are defined as economies that exhibit price distortions due to substantial governmental interference with market forces. Since 1947, there has been a presumption that prices are distorted in non-market economies where the state controls factors of production and interferes with market conditions.[2] Domestic prices are therefore thought to be unreliable in non-market economies, which leads importing countries to use alternative prices in their anti-dumping calculations.

Upon accession to the WTO, China committed to a provision that would allow WTO members to calculate dumping duties using analogue or surrogate values, constructed based on third country prices for 'normal value'.[3] Section 15(a)(ii), a subsection of the agreed protocol, was to expire after the passage of 15 years, on 11 December 2016.

On 15 December, 2016, China brought a complaint (DS516)[4] in the WTO alleging that, following the expiration of Section 15(a)(ii) on 11 December, 2016, Articles 2(1) to 2(7) of the Basic European Union (EU) Regulation on dumping investigation are inconsistent with Article I:1 of the GATT 1994, Articles 2.1 and 2.2 of the Anti-Dumping Agreement, Article VI:1 of the GATT 1994, and the second paragraph of the Second Note *Ad* to Article VI:1 of the GATT 1994 ("Second

[2] See United States Trade Representative (USTR), 'European Union – Measures Related to Price Comparison Methodologies (DS516) – Legal Interpretation – GATT 1994 Article VI:1 – Second Note Ad Article VI:1 – Practice of GATT Contracting Parties – Accessions to GATT – ADA Article 2 – and Section 15 China WTO Accession Protocol' (USTR, 13 November 2017) <https://ustr.gov/issue-areas/enforcement/dispute-settlement-proceedings/wto-dispute-settlement/pending-wto-dispute-32> accessed 13 September 2018. Here, the history of Article VI as basis for alternative price comparability in cases of non-market economies is pointed out. At page 19 specifically, it is noted that the proposed amendments after 1947 and accession protocols for non-market economies after 1947 established the incontrovertible interpretation of Article VI:1 of GATT 1947 as permitting a rejection of non-market economy prices. Also see: See WTO, *European Union — Measures Related to Price Comparison Methodologies* (10 July 2017) WT/DS516.

[3] WTO, *Accession of the People's Republic of China – Decision of 10 November 2001* (23 November 2001) WT/L/432, Section 15(a)(ii): Importing WTO Members were, subject to certain conditions, exceptionally permitted to use a methodology based on a strict comparison with domestic prices or costs in China. Also see WTO, 'Transitional Review Mechanism Pursuant to Section 18 of the Accession of the People's Republic of China – Questions from the United States' (23 October 2003) G/ADP/W/436, 5.

[4] WTO, *European Union – Measures Related to Price Comparison Methodologies* (10 July 2017) WT/DS516.

Note").[5] Furthermore, China alleges that after expiry of Section 15, Article 2(7) of the EU Basic Regulation[6] breaches the most-favoured-nation treatment required under Article I:1 of GATT 1994. China asserts that its economy does not fit the provisions of the Second Note, which is the only legal authority for non-market economy treatment.

The United States (US) has a substantial trade interest in this matter because, like the EU, it also applies alternative, surrogate methodology to its dumping margin calculations for China; it therefore joined case DS516 as a third party with a substantial trade interest in the outcome of the case. China brought a similar case against the US in case DS515/1 (concerning the expiration of Section 15(a)(ii) of the Protocol on the Accession of the People's Republic of China).[7]

III. China Continues to be Bound by Section 15 After 11 December 2016

On 6 December 2017, Ambassador Zhang Xiangchen made an opening statement before the dispute settlement panel in case DS516. He invoked *pacta sunt servanda*, the fundamental principle that 'agreements must be kept',[8] to argue that WTO Members are no longer authorised to reject Chinese prices or costs in anti-dumping investigations after 11 December 2016. Zhang argued that WTO Members were obligated by the terms of Section 15(d) of China's accession protocol to cease all 'analogue country' dumping calculation methodologies. In support of China's interpretation, Zhang cited the *Fasteners*[9] dispute, in which the Appellate Body decided that "paragraph 15(d) of China's Accession Protocol establishes that the provisions of paragraph 15(a) expire 15 years after the date of China's accession".[10] As a result, China contends that it has discharged its sole obligation under Section 15, which was to endure fifteen years of non-market

[5] WTO, *The Legal Texts: The Results of The Uruguay Round of Multilateral Trade Negotiations* (Cambridge University Press 1999); General Agreement on Tariffs and Trade (adopted 15 April 1994, entered into force 1 January 1995) 1867 UNTS 3 (hereinafter "GATT 1994"); The Second Note Ad is an exception to Articles 2.1 and 2.2 of the Anti-Dumping Agreement and Article VI:1 of the GATT 1994. It permits a Member to depart from a strict comparison with domestic (Chinese) prices if the Member satisfies the two conditions set forth in the Second Note.

[6] Regulation (EU) 2016/1036 of the European Parliament and of the Council of 8 June 2016 on protection against dumped imports from countries not members of the European Union [2016] OJ L 176/21 (hereinafter, "EU Basic Regulation").

[7] WTO, *United States – Measures Related to Price Comparison Methodologies* (12 Dec 2016) WT/DS515.

[8] Vienna Convention on the Law of Treaties (adopted 23 May 1969, entered into force 27 January 1980) 1155 UNTS 331 (hereinafter "VCLT 1969"), Article 26.

[9] WTO, *European Communities – Definitive Anti-Dumping Measures on Certain Iron or Steel Fasteners from China*, (28 July 2011) WT/DS397/AB/R, para 289.

[10] ibid. See also Weijia Rao, 'China's Market Economy Status under WTO Antidumping Law after 2016' (2013) 5(2) TCLR151, 165.

economy treatment in anti-dumping investigations in exchange for unconditional market economy treatment under all WTO agreements after 11 December 2016.

The WTO Dispute Settlement Understanding[11] provides that WTO agreements such as China's Accession Protocol are to be interpreted in accordance with "the customary rules of interpretation of public international law".[12] Among the customary rules recognised for WTO treaty interpretation is Article 31 of the 1969 Vienna Convention on the Law of Treaties (hereinafter "1969 VCLT").[13] Article 31 states that a "treaty shall be interpreted in good faith in accordance with the ordinary meaning to be given to the terms of the treaty in their context and in the light of its object and purpose".[14] It further states that "interpretation must be based above all upon the text of the treaty".[15]

With respect to the ordinary meaning of the expiry clause, China interprets the ordinary meaning of the words "in any event" in the second sentence of subsection 15(d) as establishing a categorical conclusion to non-market pricing methodology, subject to no conditions and requiring no performance on the part of China.[16] In sum, China believes its sole obligation under Section 15 is to wait fifteen years to automatically 'graduate' to market economy status after 11 December 2016.

A. Section 15 contains provisions not subject to expiry

China's proposed interpretation of Section 15 leaves only China with the benefit of its bargain, while it deprives WTO Members of the terms they negotiated in the Protocol. First, China suggests that the entire chapeau found at subsection 15(a) expired along with subsection 15(a)(ii). Nothing in the text of Section 15, however, supports this interpretation. The chapeau at subsection 15(a) is not subject to expiration dates, nor are exceptions or derogation from subsection 15(a) admitted. Countries are constrained in Section 15(a) to either use Chinese prices or costs *or* an alternative methodology that does not take into account

[11] WTO, 'Understanding on Rules and Procedures Governing the Settlement of Disputes', Annex 2 of the Marrakesh Agreement Establishing the World Trade Organization (adopted 15 April 1994, entered into force 1 January 1995) 1869 UNTS 401 (hereinafter, "Dispute Settlement Understanding").

[12] Final Act Embodying the Results of the Uruguay Round of Multilateral Trade Negotiations (15 April 1994) 1867 UNTS 14, 33 ILM 1143 Annex 2, Article 3.2; see also WTO, *India – Patent Protection for Pharmaceutical and Agricultural Chemical Products* (24 August 1998) WT/DS79/R, 47.

[13] WTO, *India – Patent Protection for Pharmaceutical and Agricultural Chemical Products* (24 August 1998) WT/DS79/R 45.

[14] 1969 VCLT, Article 31.

[15] See, for example, *Case Concerning the Territorial Dispute (Libyan Arab Jamahiriya v Chad)* (Judgment) [1994] ICJ Rep 6 at 22, para 41.

[16] WTO, 'Opening Statement by Ambassador Zhang Xiangchen as a part of the Oral Statement of China at the First Substantive Meeting of the Panel in the dispute: European Union – Measures Related to Price Comparison Methodologies (DS516)' (6 December 2017) <http://images.mofcom.gov.cn/Wto2/201712/20171213174424357.pdf> accessed 4 January 2018.

Chinese domestic prices or costs. Hence, the chapeau applies indefinitely during the pendency of China's membership or until China can prove its market economy status pursuant to subsection 15(d). No mention is made anywhere in the Protocol of modifications or revisions to the chapeau after any period of time.

Secondly, China erroneously argues that the chapeau is subordinate to its subsection; therefore, expiry of one subsection vitiates the entire chapeau. This interpretation is, however, contrary to the Appellate Body's opinion in *United States – Standards for Reformulated and Conventional Gasoline*, where the Appellate Body rejected interpretations that would "empty the chapeau of its contents" and render the remaining paragraphs of a provision meaningless.[17] In fact, Section 15(a) serves a preambular function by announcing the overall object and purpose of the Section. It lays the foundation for rules to determine whether an importing Member must use Chinese costs of production or prices to determine antidumping duties within Chinese industries under investigation or whether alternative methodology such as third country prices are appropriate. Contrary to China's construction, the provisions of Section 15 are sequenced in a logical order according to their level of importance. Section 15(a) provides the overarching framework for price comparability. This framework is *implemented in* the rules of subsections 15(a)(i) and 15(a)(ii), which are an extension of the framework, but it is erroneous to assume that subsection 15(a) derives its authority from 15(a)(ii). Rather, subsection 15(a) is the logical extension of Article VI of GATT 1994, and the Anti-Dumping Agreement referenced in the preceding paragraph.

Thirdly, while China focuses all its attention on the second sentence of subsection 15(d),[18] it overlooks the plain language of the first[19] and third[20] sentences which provide the interpretive context for the second sentence. China's proposed interpretation renders the first and third sentences of subsection 15(d) meaningless, since they condition termination of "the provisions of subparagraph (a)" on China's performance of market reforms to the satisfaction of WTO Members under their municipal laws. The use of the plural, "provisions", stands in stark contrast to the second sentence in 15(d), which references termination of subsection 15(a)(ii) only.

[17] WTO, *United States – Standards for Reformulated and Conventional Gasoline* (29 April 1996) WT/DS2/AB/R. Para 23

[18] *Accession of the People's Republic of China* (n 3), Section 15(d): "…In any event, the provisions of subparagraph (a)(ii) shall expire 15 years after the date of accession…"

[19] ibid, "…Once China has established, under the national law of the importing WTO Member, that it is a market economy, the provisions of subparagraph (a) shall be terminated provided that the importing Member's national law contains market economy criteria as of the date of accession…"

[20] ibid, "…In addition, should China establish, pursuant to the national law of the importing WTO Member, that market economy conditions prevail in a particular industry or sector, the non market economy provisions of subparagraph (a) shall no longer apply to that industry or sector."

This deliberate distinction between the various provisions of subsection 15(a) leads to the logical conclusion that if the contracting parties had intended all of the provisions of subsection 15(a) to terminate within 15 years, they would have done so explicitly.

Fourthly, a textual analysis of the first and third sentences of subsection 15(d) reveals that, to gain the benefit of market economy status, China is still bound to meet the unfulfilled promises it made upon accession. The principle of *pacta sunt servanda* requires China to either meet its obligation of showing its market economy status under WTO Members' national laws, or submit to alternative pricing valuation. Expiry of subsection 15(a)(ii) left in effect 15(a)(i), by which Chinese producers must show the prevalence of market conditions to merit a deviation from the presumption of non-market economy status. Similarly, taken as a whole, subsection 15(d) reinforces the notion that, to merit market economy treatment, China bears the burden of clearly showing it has completed the transition to market economy status to the satisfaction of Members' national laws.

China's continuing obligations are reflected in the context and ordinary meaning of subsection 15(d). The first sentence of subsection 15(d) establishes the termination of subsection 15(a) "provided that the importing Member's national law contains market economy criteria as of the date of accession". The ordinary meaning of the expression "provided that" conditions removal of subsection 15(a) on the existence of national legislation that sets forth criteria to determine market economy status. The third sentence creates an identical obligation but allows China to make an industry-specific showing. Therefore, the first and third sentences of subsection 15(d) create two obligations: (1) China must establish that it is a market economy; and (2) China must satisfy market economy status criteria under the national laws of WTO Members.

China argues that the first and third sentences of subsection 15(d) only apply before 11 December 2016; yet, there is no textual evidence for this proposition. The word "once" in the first sentence of subsection 15(d) means 'as soon as'. "Once" denotes conditionality and marks the period of time that commences after the requirement has been met by China. Therefore, the first sentence is not subject to expiry on any particular date; rather, it terminates after satisfaction of the requirement.

Similarly, the third sentence is not subject to any timeframe but is conditioned on China's performance since it reads "should China establish… that market conditions prevail". Under both the first and third sentences, China bears the burden of making an affirmative showing of its market economy status to the satisfaction of an importing WTO Member provided that the Member has clear national criteria regarding market economy status. Hence, the first sentence permits

China to establish market economy status vis-à-vis individual WTO Members *at any time* before or after the fifteen-year mark.[21]

Finally, China relies on *Fasteners* as proof that the Appellate Body has already determined that all of Section 15(a), including the chapeau, was to expire after the passage of fifteen years. In support of this argument, China relies on a statement by the Appellate Body in *Fasteners* to the effect that "[p]aragraph 15(d) of China's Accession Protocol establishes that the provisions of subsection 15(a) expire fifteen years after the date of China's accession".[22] The issue of the effect of subsection 15(d), however, was not before the Appellate Body in the *Fasteners* dispute. Subsection 15(d) provides only for the expiry of subsection 15(a)(ii). Therefore, the general statement the Appellate Body made regarding expiry in *Fasteners* is limited to subsection 15(a)(ii). The remaining provisions of subsection 15(a) and subsection 15(d) remain in force after 11 December 2016.

B. The negotiating history of section 15 confirms that china must complete its transition to merit market economy treatment

Article 32 of the 1969 VCLT allows interpreters to look to preparatory work surrounding a treaty provision and the circumstances of its inclusion in cases of ambiguity.[23] Zhang cited the negotiating history and high-level public statements made by EU and US officials to support the understanding that the non-market economy methodology would cease after fifteen years. During the course of negotiating China's Accession Protocol, China rejected inclusion of a non-market economy clause, and only acquiesced on the condition that the clause would expire fifteen years after the date of accession.[24] Therefore, China argues that the original purpose of including the fifteen-year deadline was to strike a deal whereby China would accept fifteen years of discriminatory non-market economy treatment in exchange for full termination of its non-market status as of 11 December 2016.

Where Zhang sees a 'ticking clock' burdening China with unfair discrimination, the EU and US interpret Section 15 as a major concession to China. In exchange for improved trade relations with China, WTO Members negotiated Section 15(a) to

[21] For instance, Australia and other countries have recognised market economy status, see Annex II.

[22] WTO, *European Communities – Definitive Anti-Dumping Measures on Certain Iron or Steel Fasteners from China* (28 July 2011) WT/DS397/AB/R, para 289.

[23] VCLT 1969 (n 8), Article 32: "Recourse may be had to supplementary means of interpretation, including the preparatory work of the treaty and the circumstances of its conclusion, in order to confirm the meaning resulting from the application of Article 31, or to determine the meaning when the interpretation according to Article 31: (a) leaves the meaning ambiguous or obscure; or (b) leads to a result which is manifestly absurd or unreasonable."

[24] Rao (n 10) 157.

encourage industry-level reforms during the first fifteen years of China's accession. This explains why subsection 15(a) addresses *producers* within particular industries, rather than *China*. Subsection 15(a) thus incentivised bargaining and facilitated transactional relationships between WTO Members and private *Chinese producers*. Subsection 15(a)(ii) authorised Members to *presume* Chinese producers benefited from non-market conditions without relying on any additional affirmative findings for the first fifteen years of China's membership.[25] Upon expiry of subsection 15(a)(ii), however, Members must rely on their national laws, which permit them to make industry-specific market economy determinations regarding particular Chinese producers under subsection 15(a)(i). To gain recognition of a successful nationwide or sector-specific transition, subsection 15(d) requires China to clearly show it has graduated to a market economy under Members' national laws.

Thus, the provisions of subsection 15(d) shift the burden from private producers to *China* (the sovereign) to prove its market economy status under Members' national laws. Far from conferring the windfall China argues it deserves, Section 15(d) creates a balance of power that favours WTO Members' right to make a determination regarding China's market economy status *at any time*. In other words, subsection 15(a)(ii) offered a WTO-brokered presumption that no further market economy findings were necessary to support a Member's non-market determination. After expiry of subsection 15(a)(ii), however, Members are bound to make affirmative findings of market economy status with respect to Chinese producers under subsection 15(a)(i) and with respect to China under subsection 15(d). In this way, Section 15(d) reflects the object and purpose of encouraging market reforms in China after expiry of subsection 15(a)(ii) and advancing bilateral relationships until full completion of the transition.

Also worth noting is the requirement in Section 15(d) that Members have existing market economy criteria under national laws *at the time of China's accession* ("as of the date of accession"). In this way, China was assured that its economy would be measured by unbiased standards in assessments of its market economy status even after expiry of the presumption rule in 15(a)(ii).

Zhang's interpretation disregards the continuing force of sentences one and three of subsection 15(d). Instead, he characterised both provisions as "early termination" provisions that only applied *before* 11 December 2016.[26] However, use of the words "once" and "should" in sentences one and three unambiguously denote *ongoing* requirements, conditional only on China's performance. In this

[25] See USTR (n 2) para 8.6.5.1: "One of those circumstances—rejecting Chinese prices and costs without any additional affirmative finding when the producers under investigation cannot clearly show that market economy conditions prevail in the industry producing the like product—is time-limited."

[26] WTO, 'Opening Statement by Ambassador Zhang Xiangchen' (n 16), para 11.

sense, expiry of subsection 15(a)(ii) was indeed a 'ticking clock',[27] but rather than accruing unilateral benefits to China, the 'clock' counted down the days until China fully completed its transition to market economy pursuant to Section 15(d).

Zhang's opening statement cites numerous government statements that are consistent with interpretation of subsection 15(a)(ii) as a transition methodology, which would be replaced by determinations under national laws. Zhang notes that then-US Trade Representative Charlene Barshefsky and US Senator Feinstein both noted that "the special anti-dumping methodology" was to last for fifteen years.[28] The EU Commission also remarked that the "specific procedures for dealing with cases of alleged dumping by Chinese exporters" would "remain available for up to fifteen years".[29] Yet these examples only strengthen the view that subsection 15(a)(ii) established a WTO-brokered procedure which did not affect the remaining provisions of Section 15. None of the statements reflect an understanding that subsection 15(a)(ii) would nullify determinations of market economy status under national laws. Neither do the statements show an understanding that Members would lose the fundamental, sovereign right to make market economy comparisons under Article VI of GATT 1994 after fifteen years. All of these statements are mere summations of the presumption granted under subsection 15(a)(ii). They do not address the provisions of subsection 15(d). None of the cited statements discharges China of its duty to make a showing of its market economy status under subsection 15(d). Hence, the public remarks referenced by Zhang are fully compatible with the understanding that, after 11 December 2016, competence for price comparability falls to WTO Members under their national laws.

IV. Section 15 Does Not Supplant Articles VI:1 and VI:2 of GATT 1994

A. China misapplies lex specialis

China relies on a *lex specialis* analysis of Section 15, arguing that Section 15 is a derogation from price comparability rules under Article VI and the *Anti-Dumping Agreement*.[30] *Lex specialis* is a longstanding norm of customary international jurisprudence whereby special rules override general rules ('*lex specialis derogat legi generali*').[31] The traditional reasoning for prioritising the specific over the general

[27] ibid, para 6.

[28] WTO, 'Opening Statement by Ambassador Zhang Xiangchen' (n 16), para 11.

[29] WTO, 'Overview of the Terms of China's Accession to WTO' (October 2003), para 55 <http://trade.ec.europa.eu/doclib/docs/2003/october/tradoc_111955.pdf> accessed 4 January 2018.

[30] Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade (adopted 15 April 1994, entered into force 1 January 1995) 1868 UNTS 201, Article 2.1.

[31] Peter Malanczuk, *Akehurst's Modern Introduction to International Law* (1997) 49, 56

is that particular circumstances are regulated with more clarity and certainty by special rules than by general ones.[32] Under China's *lex specialis* analysis, Section 15 adds greater precision to the general treatment given to non-market economies in Article VI:1, VI:2, Second Note of the GATT 1994 and the Anti-Dumping Agreement.[33] Hence, in China's formulation, Section 15 overrides GATT VI:1 and VI:2.

China's interpretation, however, is a misapplication of the norm of *lex specialis* exception. In fact, there are two types of *lex specialis*: (1) specialised rules that constitute exceptions to general rules; and (2) specialised rules that elaborate on the application to be given to a general rule in a particular circumstance. China erroneously argues that Section 15 operates under type (1) when in fact it has every characteristic of a rule of application under type (2). Since the rules in Section 15 are an application of Article VI and they derive their authority from Article VI, they cannot be understood as an exception.

B. Negotiating history of GATT 1994 and subsequent practice

As mentioned above, Articles VI:1 and VI:2 of GATT 1994 require "comparable prices, in the ordinary course of trade".[34] Since only market economy prices can be understood as comparable and because state control distorts the ordinary course of trade, Article VI is the legal source of alternative price methodologies. Article VI leaves price comparability at the discretion of WTO Members. Negotiating history and subsequent practice confirm this understanding. The negotiating history of the Second Note Article VI:1 confirms the longstanding practice of empowering WTO members with the authority to

[32] ILC, 'Report of the Study Group on the Fragmentation of International Law – finalized by Martti Koskenniemi' (13 April 2006) UN Doc A/CN.4/L.682, 29 <http://legal.un.org/ilc/sessions/55/pdfs/fragmentation_outline.pdf> accessed 12 January 2018.

[33] GATT 1994 (n 5); The Interpretative Second Note of Article VI from Annex I allows for alternative methodologies in countries where domestic prices are fixed by the state and the state enjoys a complete or substantially complete monopoly over trade. It states:

> Paragraph 1
>
> 1. Hidden dumping by associated houses (that is, the sale by an importer at a price below that corresponding to the price invoiced by an exporter with whom the importer is associated, and also below the price in the exporting country) constitutes a form of price dumping with respect to which the margin of dumping may be calculated on the basis of the price at which the goods are resold by the importer.
>
> 2. It is recognised that, in the case of imports from a country which has a complete or substantially complete monopoly of its trade and where all domestic prices are fixed by the State, special difficulties may exist in determining price comparability for the purposes of paragraph 1, and in such cases importing contracting parties may find it necessary to take into account the possibility that a strict comparison with domestic prices in such a country may not always be appropriate.

[34] ibid.

make market economy determination under their national laws. The Working Party Sub-Group convened in 1955, drafting a report that considered Article VI:1 to contain the "flexibility and authority to reject non-market determined prices for purposes of determining dumping".[35]

Furthermore, the US Third Party submission clarifies that the Second Note merely identifies one situation (a "substantially complete monopoly") in which it may be particularly difficult to determine price comparability. Yet, the situation described in the Second Note is not the exclusive test for determining price comparability.[36] Rather, the negotiating history of the Second Note itself reveals that Members intended price comparability to flow from Article VI:1 and VI:2. Therefore, no modification to Article VI was deemed necessary at the time the Second Note was added. This is also reflected in the plain language of Article VI which requires "comparable prices" in "the ordinary course of trade" for computation of dumping. Both the terms comparable price and in the ordinary course of trade have been historically interpreted as market economy requirements.

As early as 1957, when the GATT Secretariat undertook a largescale review of application of Article VI under WTO Members' national laws, it found that a majority of members interpreted Article VI as requiring market economy status to reach comparable prices.[37] This understanding was expressed in national legislation of Canada, South Africa, Rhodesia, the United States, Belgium, Sweden, Australia, Norway, and the United Kingdom which used expressions such as 'having a free economy', 'freely offered for sale', "market price', 'in the ordinary course of trade', 'in the open market under fully competitive conditions', and 'fair market value'. Furthermore, the Secretariat itself inserted a discussion of "the state trading problem"[38] in which it recognised that WTO members often instituted the practice of alternative methodologies to make fair price comparisons: "countries levying anti-dumping or countervailing duties on imports from State-trading economies very often rely on the price situation in comparable third markets or on consultations with the exporting country".[39]

Other accession protocols also confirm the understanding that Article VI:1 and VI:2 are the source of alternative methodologies. Poland, Romania, and Hungary were each subject to alternative anti-dumping methodologies after accession under Article VI. Therefore, Article VI itself has historically been understood to stand outside of accession agreements as the underlying legal

[35] USTR (n 2) para 4.8.3.3.

[36] ibid.

[37] ibid para 5.1–5.10.2 (citing GATT, Anti-dumping and Countervailing Duties: Secretariat Analysis of Legislation (23 October 1957) L/712 < https://www.wto.org/gatt_docs/English/SULPDF/90710019.pdf> accessed 12 January 2018).

[38] ibid 10–11.

[39] USTR (n 2) para 5.2.2.

authority for price comparability determinations in anti-dumping calculations for Members after accession.[40]

In sum, Article VI:1 of the GATT is the legal authority that underpins the Second Note, the Anti-Dumping Agreement, and Section 15 of China's Accession Protocol. The Accession Protocol merely provides a particular application of Article VI, but it does not supplant Article VI.

C. The plain language of section 15

The plain language of Section 15 indicates that the non-market economy price comparisons agreed in China's accession are derived from Articles VI:1, VI:2 GATT 1994. China argues that the Second Note is the sole and exclusive legal authority that provides for rejection of normal prices and costs. The EU points out in its brief, however, that the Anti-Dumping Agreement implements and applies the relevant provisions of the GATT 1994. The two Agreements must be interpreted and applied together in a manner that is harmonious and consistent, so as to give meaning to all provisions in both agreements.[41] The EU points out that contrary to China's argument that the Second Note provides the exclusive list of circumstances in which surrogate prices may be used in comparing export and normal prices, there are at least twenty-seven such abnormal situations arising under the provisions of Article VI:1, VI:2 of the GATT 1994, Article 2 of the Anti-Dumping Agreement, the Second Note *Ad* to Article VI, and Section 15 of China's Accession Protocol. Furthermore, the EU notes that Article 2.4 of the Anti-Dumping Agreement requires a "fair comparison" which requires comparable, market-based prices, in the ordinary course of trade, in line with the provisions of Article VI:1(a).[42] Both agreements authorise Members to use alternative price methodologies where they call for comparable prices in the ordinary course of trade.

China argues that Section 15 creates an exception to the Article VI and the Anti-Dumping Agreement. The introductory paragraph of Section 15 clearly states, however, that price comparability flows from Article VI and the Anti-Dumping Agreement.[43] The introductory paragraph makes clear that Section 15

[40] ibid.

[41] WTO, *European Union — Measures Related to Price Comparison Methodologies* (14 November 2017) WT/DA516 First Written Submission by the European Union (EU Brief) 8 <http://trade.ec.europa.eu/doclib/docs/2017/november/tradoc_156401.pdf> accessed 18 January 2018.

[42] See the Second Note Ad GATT 1994 Article VI:1, the Practice of the GATT Contracting Parties in the Application of GATT 1994 Article VI:1, 230 <https://www.wto.org/english/res_e/booksp_e/gatt_ai_e/art6_e.pdf> accessed 18 January 2018.

[43] See Accession Protocol Section 15 (n 2), introductory paragraph: Article VI of the GATT 1994, the Agreement on Implementation of Article VI of the General Agreement on Tariffs and Trade 1994 [Anti-Dumping Agreement] "the SCM Agreement shall apply in proceedings involving imports of Chinese origin into a WTO Member".

does not stand alone or in contradiction to prior agreements, but rather it must be interpreted in a manner "consistent with" the framework provided by Article VI and the Anti-Dumping Agreement. The phrase "consistent with" means compatible with. "Consistent with" does not denote an exception, but rather consonance. [44] The plain language of the introductory paragraph makes clear that the special procedures provided for in Section 15 were not intended to supplant any commitments under GATT. Therefore, China's interpretation of Section 15 as *lex specialis* that overrides price comparability assessments under GATT or the *Anti-Dumping Agreement* is inconsistent with the ordinary meaning of the introductory paragraph.

D. Application of general rules in the presence of significant ambiguities or gaps in special rules

Even in the unlikely eventuality that the panel finds Section 15 is *lex specialis* with respect to Article VI, the instant dispute shows that Section 15 has significant gaps or ambiguities; therefore, the general rules should prevail. China argues that all non-market treatment under Section 15 expires on 11 December 2016, and the EU and US argue that only Section 15(a)(ii) expires after 11 December 2016. Thus, there is sufficient ambiguity in the interpretation of Section 15 to rely on Article VI. Given the competing interpretations of Section 15, it is only logical that general rules should fill in the gaps left in the special rules.

Finally, in the unlikely event that the panel finds that the entirety of Section 15 is nullified by expiration of subsection 15(a)(ii), the panel can still find the EU Basic Regulation permissible because the Basic Regulation is in line with the EU's rights under the *Anti-Dumping Agreement*. The EU invokes its right to seek an authoritative interpretation of the provisions of a covered agreement through decision making under the WTO Agreement, pursuant to Article 3.9 the Dispute Settlement Understanding.[45] The EU will argue that Section 15 *itself* is proof that Members unanimously agreed to interpret Article VI and Article 2 of the Anti-Dumping Agreement as permitting Members to reject domestic prices and costs to make a fair comparison. Article 31(3)(a) of the Vienna Convention states that subsequent agreements such as Section 15 should be considered in interpreting a previous agreement such as Article VI and the Anti-Dumping Agreement.

Therefore, even if the panel finds that analogue methodologies may no longer be used in evaluating China's domestic prices and costs under Section 15, the EU

[44] USTR (n 2).

[45] Dispute Settlement Understanding (n 11), Article 3.9: "The provisions of this Understanding are without prejudice to the rights of Members to seek authoritative interpretation of provisions of a covered agreement through decision-making under the WTO Agreement or a covered agreement which is a Plurilateral Trade Agreement."; also see: First Written Submission by the European Union (n 41) 76.

points out that the panel may still rely on the special standard of review under *Anti-Dumping Agreement* 17.6(ii), whereby a panel may uphold *a permissible interpretation* of a treaty even where customary rules of interpretation of public international law may otherwise favour a different interpretation.[46]

V. Economic Analysis Reveals China Continues to Operate as a Non-Market Economy

A. China's steps towards a market economy status

Beyond purely legal arguments, China may also proffer economic evidence of the opening of its markets, of growth in transparency and of substantial restructuring over the past fifteen years, such that the Second *Note Ad* substantial monopoly provision should no longer apply to it. While these gains have been well documented, there is still evidence that China's transition is incomplete.

In the Third China Round Table of 2015, Yuan Yuan reviewed China's accomplishments thus far. As proof that China has embraced market forces over a command economy, Yuan pointed out that in 2013 the number of investment projects subject to government ratification was cut by 60% from the 2004.[47] Yuan indicated that China has liberalised its banking and financial sector, opened itself up to foreign investment, increased both its export and import portfolios, reduced tariffs and trade barriers, made multilateral trade agreements, declaring that an "open economic system compliant with both the WTO rules and its national situation has taken shape in China".[48]

Many statistics appear to back up Yuan's claims. Jonathan Eckart writing for World Economic Forum calls the private sector is the main driver of growth and employment in China with private sector firms producing between two-thirds and three-quarters of China's GDP.[49] China is now the world's biggest producer of concrete, steel, ships, and textiles, and has the world's largest automobile market.[50] One expert calls China a "commodities powerhouse" because it imports over half of the world's annual consumption of aluminium, and nearly half of its nickel,

[46] First Written Submission by the European Union (n 41) para 279.

[47] Yuan Yuan, 'Looking Back 14 Years after Accession: Case of China' (WTO, 2 June 2015) 5 <https://www.wto.org/english/thewto_e/acc_e/Session2YuanYuanPostAccessionLookingback-14yearafter.pdf> accessed 18 January 2018.

[48] ibid 6.

[49] Jonathan Eckart, '8 Things You Need to Know about China's Economy' (*World Economic Forum: Annual Meeting of New Champions*, 23 June 2016) para 12 <https://www.weforum.org/agenda/2016/06/8-facts-about-chinas-economy/> accessed 19 September 2018.

[50] John Ikenberry, Zhu Feng and Wang Jisi (eds), *America, China, and the Struggle for World Order: Ideas, Traditions* (Palgrave Macmillan 2015).

copper, zinc, tin and steel. China's stock market is the third largest in the world.[51] China also touts a leap in its middle class. In 2016, real urban income rose by 5.8% and a recent study found that 55% of Chinese consumers are confident that their income will continue to rise in the next five years.[52]

China may also point to a vast reduction of government control since its accession to the WTO. Since 2001, China has engaged with the US to increase economic liberalisation within the frameworks of the Joint Commission on Commerce and Trade as well as the Strategic and Economic Dialogue.[53] According to Eswar Prasad of the Brookings Institute when testifying to Congress, China has been "selectively and cautiously dismantling" government control over both the inflow and outflow of capital, resulting in a freer movement of capital both domestically and in its international portfolio.[54] The government has a stated goal of shifting "foreign exchange holdings [to] the people" (and away from the central bank).[55] As such, many holdings have moved from government entities to private households and corporations. Furthermore, China has made major currency and banking reforms as it transitions from a centrally controlled exchange rate to one more market-determined. For example, as recently as August 2015, the People's Bank of China (PBC) moved away from bank-determined opening prices on the Chinese stock market, instead pegging them to the previous day's performance at closing.[56] Furthermore, "bank deposit and lending rates have now been fully liberalised", with commercially owned banks now free to set their rates based on market forces instead of government edict.[57] Scholars speak of China as a dynamic emerging economy, stimulated not through government subsidy and regulation but by domestic consumer confidence and international investor excitement.

[51] Frank Holmes, 'How China went from Communist to Capitalist' (10 Oct 2015) *Business Insider* <http://www.businessinsider.com/how-china-went-from-communist-to-capitalist-2015-10> accessed 20 January 2018.

[52] 'China's Consumers: Still Kicking' (30 April 2016) *The Economist*, <https://www.economist.com/news/business-and-finance/21697597-free-spending-consumers-provide-comfort-troubled-economy-consumption-china-resilient> accessed 20 January 2018.

[53] US Government Accountability Office, 'US-China Trade: United States Has Secured Commitments in Key Bilateral Dialogues, but US Agency Reporting on Status Should Be Improved' (11 February 2014) <https://www.gao.gov/products/GAO-14-102> accessed 6 June 2018.

[54] Ewar Prasad, 'China's Economy and Financial Markets: Reforms and Risks' (*Brookings Institute*, 27 April 2016) <https://www.brookings.edu/testimonies/chinas-economy-and-financial-markets-reforms-and-risks/> accessed 6 June 2018.

[55] ibid.

[56] ibid.

[57] ibid.

B. China argues its market economy status is irrelevant

The EU currently uses an analogue country methodology for calculating the level of dumping for products originating in China. This means that the EU uses prices from third countries rather than use Chinese prices in normal value calculations.[58] Article 2(7) of the EU Basic Regulation lists China, Vietnam, Kazakhstan, and any non-market-economy country as being subject to a presumption of non-market economy. This non-market economy status has no end date and is applied indefinitely by the legislation. To overcome this presumption, producers must sufficiently substantiate a claim that market economy conditions prevail with respect to the manufacture and sale of their product.[59]

China argues that Article 2(7) of the EU Basic Regulation does not comport with the EU's obligation to not discriminate under Article I:1 of the GATT 1994 because it creates a presumption that Chinese products originate in a non-market economy. After 11 December 2016, China has argued that this presumption is no longer supported by the Accession Protocol. In China's view, its obligation under Section 15 was not to complete its transition to market economy, but rather to simply wait for the fifteen years to pass. In fact, Zhang asserted that the matter of China's market economy is "irrelevant" in determining whether WTO Members have the right to use alternative price methodology.[60] Therefore, in China's view, the EU's trade defence laid out in Article 2(7) is discriminatory and constitutes a breach of Section 15.

C. China's actual market economy status is essential

However, China's argument defies logic. The principal object and purpose of Section 15 is to encourage China to complete its market reforms, not to provide a loophole for China to remain a state-run economy while reaping the benefits

[58] See for example, Erdal Yalcin, Gabriel Felbermayr and Alexander Sandkamp, 'New Trade Rules for China? Opportunities and Threats for the EU' (*European Parliament's Committee on International Trade*, 29 January 2016) 12 <http://www.europarl.europa.eu/RegData/etudes/STUD/2016/535021/EXPO_STU(2016)535021_EN.pdf> accessed 25 January 2018. It is explained that EU anti-dumping margins tend to be lower than in the US because the EU uses US prices as an analogue rather than countries with similar levels of development, wages, and per capita income.

[59] EU Basic Regulation (n 6), Article 2(7)(b), provides as follows: "In anti-dumping investigations concerning imports from the People's Republic of China, Vietnam and Kazakhstan and any non-market-economy country which is a member of the WTO at the date of the initiation of the investigation, the normal value shall be determined in accordance with paragraphs 1 to 6, if it is shown, on the basis of properly substantiated claims by one or more producers subject to the investigation and in accordance with the criteria and procedures set out in point (c), that market-economy conditions prevail for this producer or producers in respect of the manufacture and sale of the like product concerned. When that is not the case, the rules set out under point (a) shall apply."

[60] See WTO, 'Opening Statement by Ambassador Zhang Xiangchen' (n 16) 13.

afforded to full market economies.[61] Therefore, nothing is quite as relevant to the question of price comparability as China's progress towards achieving market economy status.

A wealth of economic research supports the finding that China remains a state-run economy. A 2008 European Commission report on China's progress towards graduation to market economy status found that China met only one of Europe's five criteria for market economy status.[62] Similarly, "New Trade Rules for China," a 2013 report by the European Commission, found that the Chinese government continued to distort market conditions.[63] The Commission determined that China imposed restrictions on exports and imports; subsidised inputs; restricted business licenses; exercised direct state influence over corporate decision-making; lacked sound legal regimes such as property rights, bankruptcy and competition laws; and interfered with the independence of Chinese banks.[64] The World Bank issued a 2015 economic update, finding that in China the State's "direct and extensive involvement in allocating resources has no parallel in modern market economies". The World Bank subsequently withdrew the report under pressure from China. [65] Since then, in 2016, the EU Parliament issued another report finding that China has not yet 'graduated' to a market economy.[66]

More recently, in October 2017, the US Department of Commerce issued a comprehensive report concluding that China remains a non-market economy

[61] See analysis of Section 15 above.

[62] European Commission, 'Commission Staff Working Document on Progress by the People's Republic of China towards Graduation to Market Economy Status in Trade Defence Investigations (19 September 2018) 26–27 <http://trade.ec.europa.eu/doclib/docs/2009/june/tradoc_143599.pdf> accessed 13 September 2018.

[63] Lukas Gajdos and Roberto Bendini, 'Policy Briefing: Trade and Economic Relations with China 2013' (24 April 2013) Directorate-General for External Policies Policy Department <http://www.europarl.europa.eu/RegData/etudes/briefing_note/join/2013/491492/EXPO-INTA_SP%282013%29491492_EN.pdf> accessed 13 September 2018. Also see for example, Dr Markus Taube and Dr Christian Scmidknoz, 'Assessment of the Normative and Policy Framework Governing the Chinese economy and its Impact on International Competition' (*Think!Desk China Research and Consulting*, 25 June 2015) <http://www.euroalliages.com/data/1456161539THINK%21DESK%20study%20on%20MES%20to%20China%20-%20Executive%20summary.pdf> accessed 13 September 2018.

[64] Commission of the European Communities, 'Document on Progress by The People's Republic of China: Towards Graduation to Market Economy Status in Trade Defence Investigations' (19 September 2008) SEC (2008) 2503 <http://ec.europa.eu/transparency/regdoc/?Fuseaction=list&coteid=2&year=2008&number=2503&version=ALL&language=en> accessed 20 January 2018.

[65] Mark Magnier, 'World Bank Deletes Section on China from Report on Web' *The Wall Street Journal* (6 July 2015) <https://www.wsj.com/articles/world-bank-deletes-critical-passage-on-china-1435940676> accessed 20 January 2018.

[66] European Parliament, 'Resolution on China's Market Economy Status' (2016) Legislative Observatory 2016/2667 (RSP) <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2016-0223+0+DOC+XML+V0//EN> accessed 13 September 2018.

under US criteria.[67] The report found that the government of China's overall relationship with markets and the private sector results in economic distortions. Fundamentally, the Chinese Communist Party controls allocation of resources, with the state directing and channelling economic actors to meet state-planned targets. State control over the economy extends to the largest financial institutions and leading enterprises in manufacturing, energy, and infrastructure. Finally, the Chinese government strategically controls supply and demand relationships, distorting formation of exchange rates and input prices, the movement of labour, the use of land, the allocation of domestic and foreign investment, and market entry and exit.[68]

D. The EU's amended article 2(7)(b) avoids country-specific bias

Moreover, the EU has recently revamped its protocol for determining whether producer members are dumping. In November 2017, the European Parliament adopted amendments to Article 2(7)(b) of the EU Basic Regulation 2016/1036 on protection against dumped imports.[69] This overhaul in the EU's approach to non-market economy determinations removes all mention of specific countries and undertakes a less discriminatory approach that allows the EU Commission to make regular assessments of market distortions in the economies of *all* its trading partners based on the five criteria previously established as well as ILO core labour standards. The results of these assessments will be used to inform dumping complaints lodged by EU industries. In this way, the EU has pre-empted any possible gains China may make in the current litigation before the WTO because the new legislation clearly defines market economy criteria under national law and applies the criteria in a non-discriminatory manner in line with both Section 15 of the Accession Protocol and Articles I:1 and VI:1 of the GATT 1994.

Therefore, under any and all legal standards a dispute settlement panel may apply, it is unlikely that China will make a sufficient showing that market conditions prevail at the macroeconomic level. It is unlikely that any amount of evidence China presents could reverse over fifty years of historical practice by which WTO

[67] 19 US Code § 1677(18) lists six statutory criteria for designation as a non-market economy: "(i) the extent to which the currency of the foreign country is convertible into the currency of other countries; (ii) the extent to which wage rates in the foreign country are determined by free bargaining between labour and management, (iii) the extent to which joint ventures or other investments by firms of other foreign countries are permitted in the foreign country, (iv) the extent of government ownership or control of the means of production, (v) the extent of government control over the allocation of resources and over the price and output decisions of enterprises, and (vi) such other factors as the administering authority considers appropriate." <https://www.law.cornell.edu/uscode/text/19/1677> accessed 7 June 2018.

[68] See Annex 1.

[69] European Parliament Resolution of 12 May 2016 on China's Market Economy Status (12 May 2016) 2016/2667/RSP <http://www.europarl.europa.eu/sides/getdoc.do?Pubref=-//EP//TEXT+TA+P8-TA-2016-0223+0+DOC+XML+V0//EN> accessed 7 June 2018.

Members have discretion to determine market economy status. Furthermore, by most standards China's economy continues to exhibit significant price distortions caused by government control of the factors of production.[70] Rather than find discrimination in China's continued non-market economy treatment by the EU, the dispute settlement panel will likely find that the EU is entitled to make its own market economy determination under Article VI of the GATT 1994 as well as Section 15 of China's Accession Protocol.

VI. Conclusion

WTO law permits members to formulate their own definition of non-market economy within the framework of the Anti-Dumping Agreement and the Second Note to Article VI:1 of the GATT 1994. Therefore, WTO Members may continue to use alternative price comparability to calculate dumping margins with respect to Chinese products. While China proposes a reading of Section 15 that renders most of its provisions meaningless, the EU and US propose a more logical reading of Section 15 that is consistent with Article VI of the GATT 1994 and the Anti-dumping Agreement.

Furthermore, it is highly implausible that Members sought to bargain away their rights to price comparability in anti-dumping investigations, since they have enjoyed this basic right under Article VI and the Anti-dumping Agreement for more than half a century. The US and EU interpretation relies on the plain language of the text, which clearly states that only subsection 15(a)(ii) expires as of 11 December 2016. All remaining provisions of Section 15 retain their force, including China's obligation to make a clear showing of its market economy status. China argues that it had only to wait fifteen years to gain new status, yet the plain language of Section 15 show China's obligation to show it has completed its transition to a market economy. In the instant dispute China holds the key to its own jail cell. To date, China has not yet made an adequate showing of its graduation to market economy status by US and EU standards. China therefore

[70] See Annex II in Laura Puccio, 'Granting Market Economy Status to China: An analysis of WTO law and of
selected WTO members' policy' (*European Parliament Think Tank*, updated December 2015) <http://www.europarl.europa.eu/RegData/etudes/IDAN/2015/571325/EPRS_IDA(2015)571325_EN.pdf> accessed 20 January 2018.

still bears the burden of showing sufficient reforms under Members' national laws to merit market economy status.

Annex 1

Summary of US Department of Commerce Report Findings [71]

Factor 1: Despite market-oriented modifications of currency convertibility, the government retains significant restrictions on and ultimate approval power over capital account transactions, intervenes in onshore and offshore foreign exchange markets including limiting extent of price divergence between onshore and offshore markets, and does not disclose pricing criteria used to calculate parity rates their currency.

Factor 2: Despite a finding of variable wages across regions, sectors and enterprises, governmental institutions constrain free bargaining between labour and management. The state prohibits independent trade unions, refuses the legal right to strike, and unions are under control of a government-affiliated Party organ. Legal remedies for labour and wage violations are slim, and labour mobility is controlled by *hukon* (household registration), causing distortions on the supply side of the labour market.

Factor 3: Despite efforts to streamline procedures, significant barriers to foreign investment persist in the form of equity limits and local partner requirements, opaque approval and regulatory procedures, technology transfer and localisation requirements. The government, not the market, is the primary conduit or barrier for foreign investment in given sectors.

Factor 4: The government exerts significant control over ownership and means of production. State-invested enterprises (SIEs) and prevalent and their relative "economic weight" substantial compared to other major economies. The Chinese government allocates resources to, invests in and shields SIEs from market forces to achieve government, not enterprise, objectives. The CCP may appoint key personnel to corporate decision-making bodies. All land in China is the property of the state which controls rural land acquisition and monopolizes distribution of urban land-use rights.

Factor 5: The state allocates resources to influence economic outcomes by means of numerous mechanisms including, *inter alia*, investment approvals, access standards, guidance catalogues, financial supports, and quantitative restrictions. (a) Sectoral-level plans are formulated and executed with the participation of a plethora of state institutions; (b) the government exerts a high degree of control over prices which it wields to effect industrial policy objectives; and (c) the state

[71] US Department of Commerce, 'China's Status as a Non-Market Economy' (26 October 2017) A-570-053 Investigation <https://enforcement.trade.gov/download/prc-nme-status/prc-nme-review-final-103017.pdf> accessed 8 June 2018.

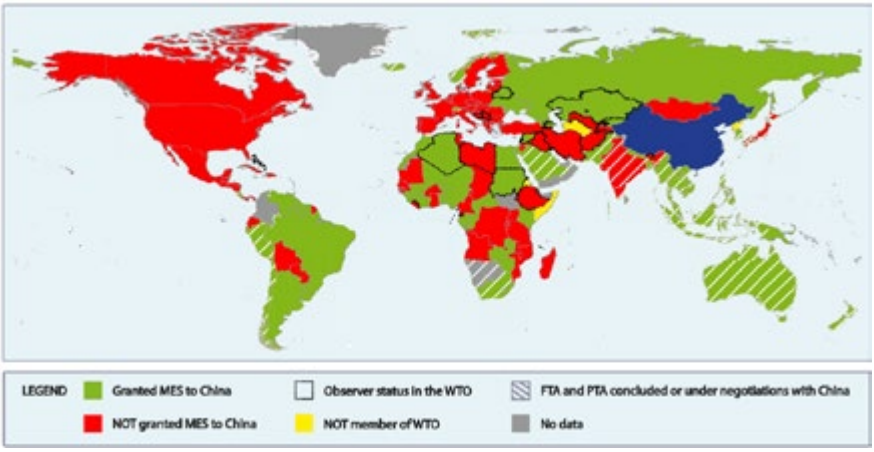
owns the largest commercial banks and oversees the majority of bank, interbank loans and even corporate bond transactions. Credit is allocated to SIEs with regard to state objectives instead of market efficiency.

Factor 6: China's legal system continues to function primarily as an instrument to achieve government and CCP-determined economic outcomes. In addition, corruption or local protectionism continues to impede the ability of firms to obtain impartial outcomes.

Annex II

China's Market Economy Status by Country (2015)[72]

Source: European Parliamentary Service



[72] Laura Puccio, 'Granting Market Economy Status to China: An analysis of WTO law and of selected WTO members' policy' (*European Parliament Think Tank*, updated December 2015) 9 <http://www.europarl.europa.eu/RegData/etudes/IDAN/2015/571325/EPRS_IDA(2015)571325_EN.pdf> accessed 20 January 2018.

A Turbulent Origin and an Uncertain Legacy: The Separation of Powers in the United States and Canada

Blake Arthur James Van Santen*

I. Introduction

Today, it is widely accepted that any state government that means to maintain the liberty of its citizens must subscribe, in one form or another, to a separation of governmental powers. Undoubtedly, the separation of powers is among the most significant and impactful political theories of modern history. It has served as an integral element of constitutional theory and a guide for institutional structure and development in states around the world for over two centuries.[1] Alongside the concept of representative government, the separation of powers has been styled "the second pillar of western political thought supporting 'constitutional' systems of government".[2]

For a theory of such moment, the meaning, and purpose, of the separation of powers are subjects of remarkable ambiguity. A look at the theory's historical underpinnings, its discussion in academic scholarship, and its invocation in jurisprudence, reveals a striking diversity in characterisations of the theory. Different historical traditions have given the concept an array of possible interpretations.[3] Today, the organisational principles associated with the theory differ greatly from

* B.A., M.A. (University of Western Ontario), J.D. (Queen's University). I would like to thank Professor Warren Newman for his encouragement and guidance in the writing of this article, as well as the editors of this journal for their indispensable assistance. I am grateful to Stephanie Walmsley for being a sounding board for my ideas and to Arthur Van Santen for his invaluable advice.

1 M J C Vile, *Constitutionalism and the Separation of Powers* (Clarendon Press 1967) 2, 7.

2 ibid 2.

3 Christoph Mollers, *The Three Branches: A Comparative Model of Separation of Powers* (OUP 2013) 40.

one country to another, and state structures ostensibly modelled on the theory's canons take a variety of forms.[4]

Though the definition of the separation of powers and the conditions purportedly required for its realisation may differ according to the time, place, and even individual in question, it would be beneficial to identify the theory's principal components to orient an analysis of the theory. Professor Maurice Vile of the University of Kent is a leading authority on the subject,[5] and helpfully articulates a strict version of the theory—dubbed the "pure doctrine"—endorsed for its accuracy even by Vile's critics.[6] Vile sets out four principles that underpin the "pure" form of the separation of powers.[7] First, a state must be divided into three branches or departments: legislature, executive, and judiciary. Secondly, all government acts must be classified as an exercise of the legislative, executive, or judicial function, and fall within the exclusive jurisdiction of the corresponding branch. Thirdly, each branch must be composed of distinct individuals; plurality of office is prohibited. Finally, if a state's institutional structure is properly constituted in accordance with the foregoing three principles, the fourth principle holds that each branch of government will necessarily act as a check on the others. In this way, each branch is confined to its respective sphere and domination by any one branch is prevented. Vile's four-faceted definition will be used as a point of reference throughout this article's analysis of the separation of powers theory.

II. Overview

This article will begin with an overview of the evolution of the separation of powers theory. This will be followed by an analysis of the theory's role in, and impact on, contemporary constitutional practice in the United States—the most renowned state-patron of the separation of powers theory. This article will then consider the theory's relationship to constitutional theory and practice in Canada, a country with historical and political traditions that diverge significantly from those of its southern neighbour. Finally, the implications of each state's distinctive

[4] ibid 43.

[5] See comments on the significance of Vile's work in: Carl J Friedrich, 'Review of *Constitutionalism and the Separation of Powers*, by M J C Vile' (1968) 73(4) The American Historical Review 1099 <https://www.jstor.org/stable/1847396> accessed 10 November 2017; H G Nicholas, 'Review of Constitutionalism and the Separation of Powers, by M J C Vile' (1967) 1(2) Journal of American Studies 294 <https://www.jstor.org/stable/27552802> accessed 10 November 2017.

[6] William Gwyn, 'Review of *Constitutionalism and the Separation of Powers*, by M J C Vile' (1969) 41(4) The Journal of Modern History 524, 526 <https://www.jstor.org/stable/446581> accessed 10 November 2017. Although Professor Gwyn asserts that Vile's analysis "ignores the sociocultural aspect of Western Constitutionalism", he nevertheless lauds Vile's definition of the "pure doctrine" as "a rather full one, including both the goal of the separation of powers and the process by which the goal is reached".

[7] Vile (n 1) 14–17.

approach to the separation of powers will be compared and contrasted, and the benefits and drawbacks of each approach will be assessed.

In the United States, the primary purpose of the theory has been to separate and balance functionally specialised governmental organisations to prevent the preponderance of any single branch of government. To this end, the state is divided into the three prescribed branches with mutually exclusive membership, in line with the first and third principles of the pure separation of powers theory. A modified version of the pure theory's second principle is also implemented: each branch is prohibited from accruing any power that does not correspond to the branch's designated function if this is not otherwise sanctioned by the narrow regime of checks and balances. However, even in its adulterated form, the principle of allocation according to function is not strictly adhered to. This may be because of pragmatic considerations or because a definitive categorisation of a specific state power is not possible.

In Canada, the executive is fused with the legislature in violation of the pure theory's prohibition on the plurality of office. This form of organisation is designed to embrace, rather than guard against, the supremacy of the legislature. Accordingly, in Canadian constitutional theory, the separation of powers doctrine does not serve the same purpose as in the United States. Instead, the doctrine serves as the nominal and protean rationale for whatever institutional arrangements happen to characterise the Canadian polity at a given time. The theory further serves to carve out boundaries which should not be crossed and as a reminder that alterations to Canada's existing state structure are, in general, to be avoided.

A comparison of the two approaches suggests that the American interpretation of the separation of powers, and the form of state organisation to which it has given rise, allows for, if it does not directly precipitate, a degree of political deadlock that is avoided, or at least less pronounced, in Canada's parliamentary system.

One conspicuous commonality between the states, relative to the separation of powers, is also revealed in this article's analysis. The first aspect of this shared quality is that both the American and Canadian constitutions explicitly eschew some of the most fundamental principles of the separation of powers theory. That both constitutions deviate, and were in fact intended to deviate, from the theory's key tenets is evident as much from the historical context of their creation as from the texts themselves. Yet, despite clear departures from separation of powers orthodoxy in both constitutions, and the drastically different structure of government organisation in each state, jurisprudence in both the United States and Canada is wont to invoke the separation of powers as each state's constitutional bedrock.

Why is this the case? This article contends that the theory's formulation in a neat, apothegmatic phrase, has given it a peculiar staying power in the ethos

of constitutionalism. The apparent appeal of the phrase "separation of powers" has, however, led to its indiscriminate application. The theory is thus an unsettled one—its principles are susceptible to different treatment, and a given principle may be interpreted either as a non-derogable rule, a flexible guideline, or ignored altogether. In practice, the doctrine is invoked in discrete situations to lend a veneer of legitimacy to laws and decisions that tend toward a separation of state powers along functional lines. However, because the theory is so nebulous, its normative value is extremely limited. It is hardly possible for state organisation to be guided by a theory espousing principles of an unknown quantity. Compliance with the separation of powers is therefore often elusive, and it is uncertain when courts will conclude that the separation of powers has been unacceptably violated and order remedial action.

III. Evolution of the Separation of Powers

The separation of powers theory has its roots in the theory of mixed government.[8] This latter theory is premised on the participation of each major societal order in the core aspects of government.[9] It was grounded in recognition of the need for each societal element to be able to protect its own distinct interests. Proponents believed this was achieved by assigning each order responsibility for an assortment of state activities, which would have the effect of preventing the dominance of any single order in the broader governance of the state.[10] This theory is of ancient pedigree: Plato's Laws, Aristotle's Politics and Polybius' Histories evince the principle's influence on the governmental structures of ancient Athens and Rome.[11] The mixture described by Plato was of monarchy and democracy, while Aristotle was concerned with democracy and oligarchy. A threefold mixture of monarchical, aristocratic, and democratic elements characterised the Roman polity depicted by Polybius. Under the Republic, each Roman order participated in the state's various functions and was thereby able to exert a restraining influence on its counterparts. Of this system, Cicero remarked, "such a government insures

[8] Vile (n 1) 3; Mollers (n 3) 46; Bruce Ackerman, *We the People: Foundations* (Harvard University Press 1991) 217–218.

[9] Vile (n 1) 33.

[10] Ackerman (n 8) 217–218.

[11] William Gwyn, *The Meaning of the Separation of Powers* (Tulane University 1965) 24.

[sic] at once an element of equality, without which the people can hardly be free, and an element of strength".[12]

A. Separating the powers: the English civil war

The tripartite organisation of the Roman system was not, broadly speaking, dissimilar from the structure of the English government on the eve of that country's Civil War, from which event the separation of powers first emerged as a distinct theory of government.[13] In his response to the demands of the Long Parliament embodied in the Nineteen Propositions of 1642, King Charles I extolled the benign effects that flowed from the kingdom's existing mixture of monarchy, aristocracy, and democracy:

> The experience and wisdom of your ancestors hath so moulded this [government] out of a mixture of these, as to give this kingdom (as far as human prudence can provide) the conveniences of all three, without the inconveniences of any one, as long as the balance hangs even between the three estates, and they run jointly on in their proper channel.[14]

The king's response was grounded in political philosophy stretching back to Aristotle that had conceived of state organisation in terms of tasks, such as agricultural, military, and financial, rather than in terms of functions.[15] However, the seventeenth-century contest between the Crown and Parliament which culminated in the Civil War had brought two distinct functions of government into sharp relief: legislating on one hand, and executing the law on the other.[16] To the Parliamentary faction, the importance of these two broad, though originally ill-defined,[17] functions was not taxonomic, but normative. Their separation was necessary for the achievement of a desired end: defence of the Englishman's famed

[12] Marcus Tullius Cicero, *On the Commonwealth* (James E G Zetzel tr, CUP 1999) book 1, ch 45.

[13] Gwyn (n 11) 37; Vile (n 1) 3.

[14] Charles I, 'Propositions Made by Both Houses of Parliament… with His Majesties Answer Thereunto' (1642) <http://oll.libertyfund.org/pages/1642-propositions-made-by-parliament-and-charles-i-s-answer> accessed 13 November 2017.

[15] Vile (n 1) 27.

[16] ibid 21, 25.

[17] Gwyn (n 11) 28–30. The distinction grew out of an earlier dichotomy between legislation (including taxation) and the "functions of government", or the royal prerogative, which included more than just the execution of laws. In the day-to-day management of the country, it was accepted that the king was capable of exercising legislative as well as executive and judicial functions. The ambiguity of the limits of these two broader functions was a significant contributing factor in bringing about the Civil War.

civil liberty from the caprice of arbitrary government.[18] For, though the legislature was capable of both enacting limits on the Crown's activities and holding the Crown's agents accountable for abuses, these safeguards were jeopardised by the Crown's participation in legislation.[19] The prevailing system of mixed government allotted tasks to the Commons, the Lords, and the Crown, vouchsafing for each a role in the broader function of legislating. The Crown was thus armed with the means of frustrating the restraining impulses of the Houses of Parliament, while the latter obstinately refused to exercise their tax-raising powers to finance, among other things, Charles' military schemes.[20]

The resulting impasse between Royalist and Parliamentary factions caused disillusionment with the efficacy of the mixed government model.[21] The turning point in the development of the separation of powers theory came with the evolution of the Parliamentary position from advocating Parliament's dominance in legislating, to demanding the king's complete exclusion from the legislature. The monarch was deprived even of his "negative voice", or veto, and was wholly confined to carrying out, or executing, the laws enacted by Parliament.[22] Though the English system had not yet recognised a separate judicial function of government,[23] all four elements of Vile's pure theory of the separation of powers were otherwise realised at this point. Charles, however, ever animated by notions of the divine right of kings,[24] did not meekly suffer the diminution of what he saw as his royal prerogative. Civil War convulsed England from 1642 to 1646, and again briefly in 1648.[25] Victory ultimately fell to the Parliamentarians. On 31 January 1649, even as he mounted the scaffold outside the Banqueting House in Whitehall where he was to be executed, Charles met his fate decrying the inroads that Parliament had carved into royal privilege.[26]

Yet the constitutional situation remained volatile. Even before Charles' departure, the burgeoning power of Parliament had led to excesses as tyrannical as those of the monarchy that had precipitated civil upheaval in the first place. After Charles fell into Parliament's custody in 1647, the prospect gradually arose of a rapprochement between Crown and Parliament. To pre-empt any such

[18] ibid 8–9.

[19] ibid 35.

[20] Trevor Royle, *Civil War: The Wars of the Three Kingdoms, 1638–1660* (Little, Brown & Company 2004) 20–25.

[21] Vile (n 1) 39.

[22] ibid 41–43.

[23] ibid 27. The idea of a 'judicial power' vested in the House of Lords, did come from the Civil War, but was a peripheral development, and never matured into a freestanding principle.

[24] Royle (n 20) 25.

[25] Patrick Little, *The English Civil Wars* (Oneworld Publications 2014) 150.

[26] Dame Cicely Veronica Wedgwood, *The Trial of Charles I* (World Books 1964) 191–192.

development, the New Model Army, aligned with the more radical elements of Parliament, purged Parliament of those deemed too royalist or Presbyterian in their sympathies.[27] The remaining members—who comprised what came to be known as the "Rump Parliament"—rode roughshod over what remained of the constitutional order. The Commons passed an ordinance to have the king tried for treason, and, when this was opposed by the House of Lords, whose concurrence was constitutionally required for any such act, declared itself the "supreme power" in the nation capable of unilaterally passing whatsoever legislation it pleased.[28] In this manner, the trial and execution of the monarch proceeded. Within a week, the Rump Parliament approved a motion to abolish the House of Lords, that "useless and dangerous"[29] body that had proved so unamenable to its regicidal designs, and whose existence as a check on the power of the Commons was now deemed inexpedient.[30] The ruthlessness of the Rump Parliament continued unabated; factions within the self-perpetuating body menaced opponents with special committees that both administered Parliament's laws and summarily adjudicated breaches.[31] Dissatisfaction spread over Parliament's continuous accrual of power and assumption of both judicial and executive roles.[32] Parliament's refusal to heed calls for reform led, in 1653, to its dissolution by Oliver Cromwell, who re-established executive dominance by declaring a Protectorate and fashioning himself Lord Protector.[33]

By the time of the Restoration, following the death of Cromwell in 1658 and the collapse of his son Richard's short-lived Protectorate, the Royalist Sir Charles Dallison could cogently sum up the lessons of the preceding decades of political turbulence: "Whilst the Supremacy, the power to judge the law, and the authority to make new laws, are kept in several hands, the known law is preserved, but united it is vanished, instantly thereupon, and arbitrary and tyrannical power is introduced."[34] English political theorists thus came away from the Civil War with a well-founded fear of arbitrary rule, whether by King or Parliament, and a novel conception of governmental organisation to preclude these eventualities: the separation of state activities into comprehensive, mutually exclusive categories

[27] Royle (n 20) 480, 484.
[28] Samuel Rawson Gardiner, *History of the Great Civil War, 1642–1649: 1647–1649*, vol 4 (Longmans, Green & Co 1893) 290.
[29] C H Firth and R S Rait (eds) *Acts and Ordinances of the Interregnum, 1642–1660* (His Majesty's Stationery Office 1911) 24.
[30] Royle (n 20) 505–506.
[31] Vile (n 1) 43.
[32] Gwyn (n 11) 33–34, 37.
[33] ibid 33–34.
[34] Vile (n 1) 46.

according to function. The trick appeared to be keeping the state organs confined to their designated function.

B. Taming the legislature with checks & balances: locke and montesquieu

Appearing in 1689, from amongst the eddies of constitutional theory swirling about in the wake of the Civil War, John Locke's Second Treatise of Government had a major impact on the development of the separation of powers theory.[35] In that work, Locke set out to reconcile the nascent theory with the supremacy of Parliament. Locke explained the natural supremacy of the legislative function by virtue of it preceding the executive function, and setting the laws by which executive power may be exercised: "[W]hat can give laws to another, must needs be superior to him."[36] The articulation of the supremacy of law was a crowning achievement of the English Civil War and remains a central tenet of democratic political theory.[37] But the danger of legislative despotism had drawn attention to the need for a balancing of the constitution. Locke, drawing on the older theory of mixed government, advocated a legislative role for the executive that would allow it to check the excesses of the legislature. The executive veto was reintroduced to balance the legislative and executive powers and ensure that the chief executive (the king) "is no more subordinate than he himself shall think fit."[38]

During the Civil War, the categorical separation of state functions, in conformity with the second principle of the pure theory of the separation of powers, had laid bare the dangers of legislative supremacy. Specifically, the English experience had exposed the futility of the theory's fourth principle: the conceptual safeguards inherent in creating multiple autonomous branches with their own institutional interests could not ensure the proper allocation of state functions and the maintenance of each branch within its defined sphere.[39] The rampant self-aggrandisement of the Long Parliament attested to the theory's deficiency in this regard. Locke's solution was an intermingling of state functions between the branches to serve as checks. This violation of the second principle of the separation of powers theory was found necessary, in light of the impotence of

[35] ibid 58.
[36] John Locke, *Second Treatise of Government* (Aziloth Books 2013) 79.
[37] Cooper v Canada [1996] 3 SCR 854 [23]; *Reference re Remuneration of Judges of the Provincial Court of Prince Edward Island* [1997] 3 SCR 3 [139].
[38] Locke (n 36) 80.
[39] Vile (n 1) 139–140, 146, 148.

the fourth principle, to keep the powers, if no longer wholly separate, then at least substantially so.

The credit for founding the separation of powers theory generally falls to Locke's intellectual disciple, Charles-Louis de Secondat, Baron de La Brède et de Montesquieu. The French writer's celebrated 1748 publication, The Spirit of the Laws, was a scientific study of governments throughout history that sought to demonstrate a causal connection between the nature and form of a state's government and the laws of that state.[40] The book's most influential chapter, credited with establishing the separation of powers doctrine in earnest, was, ostensibly, an analysis of the English Constitution, and drew heavily on the ideas of contemporary English writers and political theorists including Locke.[41]

While the separation of powers theory may not, in fact, have originated with Montesquieu, he was responsible for its apotheosis and enduring status as a universal constitutional precept. To an unprecedented degree, Montesquieu emphasised the theory as the essential element of any constitution that had political liberty as its aim.[42] Significantly, Montesquieu was also the first to clearly demarcate the "power of judging" as an independent function.[43] Conceptually, Montesquieu also removed the king from the legislature, where he continued to occupy a role in English theories by virtue of his veto over legislation.[44] Montesquieu was thus the first to express the pure theory's first principle in its modern formulation, requiring separation of the state into three distinct branches: legislative, executive, and judicial. However, Montesquieu also proposed a Lockean intermingling of functions to provide the executive branch with an active means of defence against the legislature. Montesquieu advised that, in addition to a veto, the executive ought to possess the power to convene and regulate the duration of meetings of the legislature, because "if the executive does not have the right to check the enterprises of the legislative body, the latter will be despotic… since it will be able to give to itself all the power it can imagine."[45] Thus, Montesquieu's articulation of the separation of powers theory, like Locke's, prescribed checks and balances

[40] ibid 76.

[41] ibid 58, 85.

[42] Sir William Ivor Jennings, *The Law and the Constitution* (University of London Press 1960) 20–21.

[43] Vile (n 1) 104. However, in decreeing that judgments should always be "a precise text of the law" to avoid the uncertainty inherent in judicial "opinions", Montesquieu deemed the judiciary to be of far less significance than it would come to assume as the theory matured. See: Montesquieu, *The Spirit of the Laws* (Anne Cohler, Basia Miller, and Harold Stone trs, CUP 1989) 158.

[44] Montesquieu (n 43) 164.

[45] ibid 162.

reminiscent of the theory of mixed government. Ironically, it was this very theory that the separation of powers doctrine had been explicitly designed to supersede.

On the surface it appeared that, for Montesquieu, the definitive characteristic of the English Constitution, responsible for that country's then unrivalled political liberty, was the separation of the three fundamental powers of government. The true subject of Montesquieu's fancy, however, is a matter of some debate. Noting the English monarch's role in the legislature and in judicial appointments, as well as the Lords' prerogative as the supreme court of appeal, James Madison asserted that, "on the slightest view of the British constitution, we must perceive that the legislative, executive, and judicial departments are by no means totally separate and distinct from each other."[46] A century later, Albert Venn Dicey, the eminent British constitutional theorist, similarly commented on the incongruence between the actual relationships between branches of the English government and Montesquieu's caricature of them. Dicey concluded that Montesquieu "misunderstood the principles and practice of the English Constitution on this point."[47] In fact, it has been contended that Montesquieu was not describing the English Constitution at all. Rather, Montesquieu was describing the ideal constitution for the prevention of tyranny. As his point of reference, Montesquieu chose the constitution of a nation which, in the preceding century, had overthrown two despotic monarchs, and which contrasted sharply with his native France, then labouring under the yoke of monarchical absolutism.[48] For many, then, Montesquieu's famous chapter was not a description of the English Constitution, but a prescription for a constitution at once conducive to liberty and repugnant to tyranny.[49]

C. Trial and error: the American revolution and the founding decade

It was therefore natural that, in the latter half of the eighteenth century, discontented and democratically-minded subjects in the British colonies in America couched criticism of their government in the language of Locke, and increasingly, after the publication of The Spirit of the Laws, in terms of the separation of powers.[50] Under Britain's colonial administration, governors administered the colonies in a more or less arbitrary fashion, free from the restraints placed on the

[46] James Madison, Alexander Hamilton, and John Jay, *The Federalist Papers* (Penguin Books 1987) No. 47, 303.
[47] Albert Venn Dicey, *Introduction to the Study of the Law of the Constitution* (St Martin's Press 1961) 338.
[48] Jennings (n 42) 21–23.
[49] Vile (n 1) 77.
[50] ibid 122, 127.

executive back in England, and in no way beholden to the representative colonial legislatures. Thoroughly aristocratic governor's councils dominated all state functions, forming the legislative Upper Houses, governors' advisory councils, and the colonies' supreme courts.[51]

After throwing off the bonds of imperialism, the newly independent American States drafted constitutions based on the separation of powers.[52] Initially, a strict separation of powers was much in vogue. To varying degrees, reform-minded framers of the early state constitutions rejected the concept of checks and balances as a loathsome vestige of the antidemocratic, class-based system of mixed government.[53] The pure theory of the separation of powers thus experienced a renaissance. The revolution had been galvanised by the idea that all state authority emanated from the people. Because the people directly delegated their authority to elected representatives in the distinct branches of government, whose accountability was maintained by periodic elections, it was thought unnecessary and undesirable for the legitimately held authority of each department to be subject to interference from the other branches.[54] Accordingly, aside from electoral sanction, the early state constitutions placed their faith in the fourth principle of the pure theory of the separation of powers and relied exclusively on the theory's intrinsic conceptual safeguards to maintain the branches of government within their respective bounds.[55]

It is somewhat surprising that the pure model of the separation of powers experienced the revival that it did. Strict separation had undergone an abortive experiment in England in the previous century, after which Montesquieu, the "oracle" of the American Revolution,[56] had warned of the dangers of a rigid separation, hence Dicey's observation that Montesquieu's doctrine was either misunderstood, exaggerated, or misapplied by its revolutionary proponents of the eighteenth century.[57]

In a sequence of events reminiscent of the English experience following the Civil War, the American state legislatures, bereft of positive restraints, quickly permeated all spheres of government activity. Those bodies soon accumulated a disproportionate degree of power in their hands, which was not infrequently used

[51] ibid 127; Robert F Williams, 'Evolving State Legislative and Executive Power in the Founding Decade' [1988] 496 The Annals of the American Academy of Political and Social Science 43, 44 <http://www.jstor.org/stable/1046317> accessed 18 November 2017.

[52] Vile (n 1) 135; Williams (n 51) 44.

[53] Malcolm P Sharp, 'The Classical American Doctrine of "The Separation of Powers"' [1935] 2(3) The University of Chicago Law Review 385, 396 <https://doi.org/10.2307/1596321> accessed 10 January 2018; Vile (n 1) 139, 141; Williams (n 51) 43, 45.

[54] Vile (n 1) 141.

[55] ibid 139–40, 146, 148; Williams (n 51) 45–46.

[56] Madison, Hamilton, and Jay (n 46) No. 47, 303.

[57] Dicey (n 47) 338.

in an arbitrary manner.[58] As James Madison observed at the time, in an essay that would later form part of The Federalist Papers, "the legislative department is everywhere extending the sphere of its activity and drawing all power into its impetuous vortex".[59] It was now the turn of the Americans to address the reality that, as Madison put it, "in a republican government, the legislative authority necessarily predominates".[60]

Not everyone in the nascent Republic had been under the same illusions about the wisdom of reviving the pure theory, and many had warned against it.[61] Madison, in particular, disparaged the illusory divisions between the branches of government—the "parchment barriers"—relied upon in the state constitutions.[62] Preoccupied with ensuring the subjugation of the executive following their escape from monarchical oppression and despotic colonial administration, Madison observed that the framers of the new republican constitutions "seem never to have recollected the danger from legislative usurpations, which, by assembling all power in the same hands, must lead to the same tyranny as is threatened by executive usurpations".[63] Madison elucidated the lesson that America had been forced to learn for itself: "[T]he mere demarcation on parchment of the constitutional limits of the several departments is not a sufficient guard against those encroachments which lead to a tyrannical concentration of all the powers of government in the same hands."[64]

As in England after the odious reign of the Long Parliament, there followed in America a backlash against the extreme form of the separation of powers. While the theory continued to underpin state constitutions, it was supplemented with checks and balances—mechanisms that had been shouted down as monarchical derogations from the separation of powers in the revolutionary pique of 1776.[65] Madison noted of the continued presence in state constitutions of language suggestive of the pure theory that, "notwithstanding the emphatical and, in some instances, the unqualified terms in which this axiom has been laid down, there is not a single instance in which the several departments of power have been kept absolutely separate and distinct."[66] With this development, the second principle of

[58] Vile (n 1) 143–144.
[59] Madison, Hamilton, and Jay (n 46) No. 48, 309.
[60] ibid No. 51, 320.
[61] Vile (n 1) 147.
[62] Madison, Hamilton, and Jay (n 46) No. 48, 309.
[63] ibid No. 48, 309.
[64] ibid No. 48, 312.
[65] Vile (n 1) 148.
[66] Madison, Hamilton, and Jay (n 46) No. 47, 305.

the pure theory, stipulating a rigid separation of powers according to function, was banished from American constitutional practice.

After experience with the early state constitutions had underscored the sagacity of active checks on the power of each branch of government, the atmosphere was conducive to their inclusion in the Federal Constitution of 1787.[67] To weaken the legislative branch, Madison secured agreement for a divided, bicameral legislature.[68] A plethora of cross-functional roles further ensured the interdependence of the executive and legislative branches. The chief executive, the President, was given a veto over the legislature, though one subject to override by two thirds of Congress. The President received the power to appoint his own magistrates as well as judges, though subject to confirmation by the Senate. The power to negotiate treaties was vested in the President, again qualified by the requirement of senatorial confirmation. The President was accorded the position of Commander-in-Chief of the armed forces, but Congress retained the power to declare war.[69] It must be noted, however, that what has been styled "the greatest of these checks and balances"[70] came more than a decade after the promulgation of the Federal Constitution. In 1803, Justice Marshal's famous dictum in Marbury v Madison established the judicial prerogative to review both legislative and administrative action for constitutional compliance. This added another potent check on the improper exercise of power to those set out in the Federal Constitution. Specifically, Justice Marshal erected a formidable judicial barrier against the perennial danger of legislative tyranny, should the executive veto prove insufficient.[71]

The American system was therefore contrived such that oppressive state measures required, in most instances, the cooperation of at least two branches of government.[72] Each of those branches was to exercise its own function as independently as possible. However, to maintain its independence, each branch required the ability to interfere in the functions of other, overly ambitious branches. Experience had shown that, "unless these departments be so far connected and blended as to give each a constitutional control over the others, the degree of separation which the maxim required, as essential to a free government, can never in practice be duly maintained".[73] The emphasis, though, remained on the separation of powers, because, to the extent that each branch had a role in the functions of the others, it was not aimed at fusing the branches, but erecting

[67] Vile (n 1) 148; Williams (n 51) 53.
[68] Madison, Hamilton, and Jay (n 46) No. 51, 320.
[69] Vile (n 1) 156.
[70] Jennings (n 42) 27.
[71] *Marbury v Madison* 5 US (1 Cranch) 137 (1803), [177]–[178].
[72] Laurence H Tribe, *American Constitutional Law* (Foundation Press 1978) 16.
[73] Madison, Hamilton, and Jay (n 46) No. 48, 308.

effective barriers between them.[74] An effective separation of the branches, not their fusion, was still regarded as the bulwark against the tyranny of any one branch.

IV. The Separation of Powers in the United States

A. Flexible separation: by design and necessity

The historical evolution of the separation of powers theory in America demonstrates that adherence to the theory in the United States is qualified to a considerable extent. The pure theory's first and third principles are dutifully observed in the organisation of state institutions into executive, legislative, and judicial branches, with mutually exclusive membership. The theory's American modifications are most evident in relation to the second principle's injunction to allocate state powers according to function. Derogation from the pure theory's second principle had, however, been inevitable, even if it had not been considered the most effective means of suppressing tyranny. In his concurring opinion in the United States Supreme Court's 1986 decision in Bowsher v Synar, Justice Stevens explained that "one reason that the exercise of legislative, executive, and judicial powers cannot be categorically distributed among three mutually exclusive branches of government is that governmental power cannot always be readily characterised with only one of those three labels". Justice Stevens went on to observe that, "as our cases demonstrate, a particular function, like a chameleon, will often take on the aspect of the office to which it is assigned".[75] In Youngstown Sheet & Tube Co v Sawyer, Justice Frankfurter, with more concision and less colour, asserted that, "the content of the three authorities of government is not to be derived from an abstract analysis. The areas are partly interacting, not wholly disjointed".[76]

As these statements by Supreme Court Justices make clear, there is no definitive test for determining which category a particular state power properly falls under. Thus, even if the Framers had not consciously rejected a strict separation according to function in favour of one involving checks and balances, pursuing a rigid separation would have led to arbitrary results, and would likely have proved unworkable. Consider the 1983 case of Immigration and Naturalization Service v Chadha,[77] where the Supreme Court grappled with the constitutionality of a congressional veto over the Attorney General's decision to suspend the deportation of an illegal immigrant. As the prominent American constitutional law scholar Laurence Tribe noted, the decision on deportation could have been branded

[74] Vile (n 1) 153.
[75] *Bowsher v Synar* 478 US 714 (1986) [749].
[76] *Youngstown Sheet & Tube Co v Sawyer* 343 US 579 (1952) [610].
[77] *Immigration and Naturalization Service v Chadha* 462 US 919 (1983).

legislative, executive, or judicial, depending on whether it was taken by the House of Representatives, the Attorney General, or an administrative tribunal.[78] Recourse to the pure theory's second principle of allocation based on function is of little use in determining which branch of state properly has jurisdiction over such a decision.

B. Lingering division on the second principle: guideline or dogma?

Ambiguity has nevertheless persisted concerning the status of the pure theory's second principle in American constitutional practice. Even after the promulgation of the Madisonian Constitution, the separation of powers continued to be invoked as if it mandated rigid separation. This can be seen in Supreme Court decisions of the late nineteenth century, such as Kilbourn v Thompson, which adopted the extreme view that each branch must "be limited to the exercise of the powers appropriate to its own department and no other."[79] Similarly absolutist interpretations continue to find expression in modern jurisprudence. In the 1988 Supreme Court decision in Morrison v Olson, Justice Scalia prefaced his dissent with the archetypical formulation of separation of powers orthodoxy, found in Part the First, Article XXX, of the Massachusetts Constitution of 1780:

> In the government of this Commonwealth, the legislative department shall never exercise the executive and judicial powers, or either of them: The executive shall never exercise the legislative and judicial powers, or either of them: The judicial shall never exercise the legislative and executive powers, or either of them: to the end it may be a government of laws, and not of men.[80]

Of this exact provision Madison had observed that, "[i]n the very constitution to which it is prefixed, a partial mixture of powers has been admitted". In reality, Madison continued, that constitution's interdiction "goes no farther than to prohibit any one of the entire departments from exercising the powers of another department".[81] Yet, after reciting this tract, Scalia proceeded to read into the Federal Constitution's vestment clauses exactly what American history and the Framers warned against: a categorical, unyielding separation of state powers along functional lines. Reproducing the second of the Constitution's three vestment clauses, which states that "[t]he executive Power shall be vested in a President of the United States of America", Scalia declared in no uncertain terms that, "this

[78] Laurence H Tribe, *American Constitutional Law* (3rd edn, Foundation Press 2000) 137.
[79] *Kilbourn v Thompson* 103 US 168 (1880) [103].
[80] *Morrison v Olson* 487 US 654 (1988) [697].
[81] Madison, Hamilton, and Jay (n 46) No. 47, 305–306.

does not mean some of the executive power, but all of the executive power".[82] On this footing, Scalia excoriated the majority's decision to uphold the validity of the impugned legislation providing for independent investigation into executive misconduct. Governmental investigation and prosecution of crimes, argued Scalia, is a "quintessentially executive function" and depriving the President of exclusive control over that power "is enough to invalidate the statute".[83]

Justice Scalia's interpretation of the separation of powers ignores the lessons of history, the exigencies of indeterminate categorisation, and the character of the Constitution itself. The reality is that the Madisonian version of the theory that won out in the Federal Constitution admits of a more flexible separation. No statement comparable to that of Article XXX of the 1780 Massachusetts Constitution can be found in the Federal Constitution—a similar provision was deliberately rejected in the early stages of its drafting.[84] However, lexical semantics and the nebulous nature of the separation of powers theory have permitted the debate over rigid separation to continue, even after the issue was definitively decided in the Federal Constitution.

The Constitution's rejection of a rigid separation was not interpreted uniformly, even amongst the Framers. Certainly, some of the Framers were cognisant of the compromise they had struck between ideological orthodoxy and practicability, and felt that its Americanisation had not emasculated, but improved the separation of powers theory. As Alexander Hamilton observed:

> [T]he separation of powers has been shown to be entirely compatible with a partial intermixture of those departments for special purposes, preserving them, in the main, distinct and unconnected. This partial intermixture is… not only proper but necessary to the mutual defense of the several members of the government against each other.[85]

Others, however, and Madison in particular, did not believe that a compromise had been struck at all. Madison considered the doctrine, properly understood, never to have required a rigid separation of powers according to function. Allegations that the checks and balances of the Federal Constitution violated the separation of powers were "warranted neither by the real meaning annexed to that maxim by its author [Montesquieu] nor by the sense in which it has hitherto been

[82] Morrison (n 80) [705] (emphasis in original).
[83] ibid [705]–[706].
[84] Tribe (n 78) 128.
[85] Madison, Hamilton, and Jay (n 46) No. 66, 384.

understood in America".[86] Madison believed the theory only required that "the powers properly belonging to one of the departments ought not to be directly and completely administered by either of the other departments" and that "none of them ought to possess, directly or indirectly, an overruling influence over the others in the administration of their respective powers".[87]

Madison was committing a fallacy by claiming to adhere to the "real meaning" of the separation of powers while advocating for checks and balances. As Tribe observes, "it is a misnomer of intellectual history [that] 'separation of powers' is often used as a shorthand phrase for the complex system of checks and balances created by the Constitution which in fact mingle the different types of governmental power".[88] Madison's conceptual obfuscation is part of the reason modern constitutional scholars assert that "the separation of powers may well be the most misunderstood part of the Constitution; certainly misunderstandings of it date from the moment it was brought into being in the document".[89] The ambiguity created by the use of the phrase "separation of powers" to describe a constitution characterised by checks and balances has allowed interpretations such as those found in Kilbourn v Thompson and the Morrison v Olson dissent to persist. Advocates of a rigid separation can justifiably claim that such a principle follows directly from basic separation of powers theory. After all, does not the separation of powers stand for just that, a separation of powers? However, it is untrue to claim that the separation of powers model that guided Madison and his fellow Framers in drafting the Constitution called for a rigid separation.

C. Tyranny, not pedantry: the true criterion for application

The issue with interpretations advocating a strict separation of powers along functional lines, in addition to their impracticability and reliance on historically and textually inaccurate readings of the Constitution, is that they place undue emphasis on adherence to the theory's principles at the expense of the theory's ultimate aim. The purpose of the theory had always been to prevent a "tyrannical concentration of all the powers of government in the same hands."[90] The Founding Fathers were never concerned with ideological orthodoxy for its own sake. As Gwyn notes, the Framers were more intent on laying the foundations of stable government "than with creating a system of government based on the

[86] ibid No. 47, 308.
[87] ibid No. 48, 308.
[88] Tribe, *American Constitutional Law* (3rd edn) (n 78) 137.
[89] Robert A Goldwin and Art Kaufman, *Separation of Powers—Does it Still Work?* (American Enterprise Institute 1986) 138.
[90] Madison, Hamilton, and Jay (n 46) No. 48, 312.

abstract maxims of political philosophers".[91] Accordingly, the application of the separation of powers theory in the departmental allocation of state powers does not hinge on the dictates of an abstract principle of state organisation. The true criterion for the theory's invocation is whether there is a danger of one branch's power being augmented or diminished to the point where either that branch's own independence or that of another branch is threatened.

This explains both the inconsistent application of the theory's principles and why the rigidity of Scalia's interpretation is unwarranted. Of course, a more rigid approach is taken "where the constitution by explicit text commits the power at issue to the exclusive control" of a specific branch, in which case the Supreme Court has "refused to tolerate any intrusion."[92] Generally, however, the theory is applied only where necessary to achieve its objective. Thus, the Court has "upheld statutory provisions that to some degree commingle the functions of the Branches, but that pose no danger of either aggrandizement or encroachment".[93] At the same time, "the Court has not hesitated to strike down provisions of law that either accrete to a single branch powers more appropriately diffused among separate branches or that undermine the authority and independence of one or another coordinate branch."[94]

In line with this flexible, principled approach, the Supreme Court in Morrison v Olson allowed the removal from the President's purview of a state power it recognised to be executive in nature. Though the President was denied control over both the independent counsel and his investigation into executive misconduct, neither were these permitted to come under the control of Congress or the judiciary. The Court therefore reasoned that the separation of powers was not unduly affected by this failure to allocate a state power along functional lines given that the executive was not significantly impaired and no branch increased its power at the expense of any other.[95] Conversely, in instances perceived to be more portentous, the Court has not hesitated to intervene under the auspices of upholding the separation of powers. In Immigration and Naturalization Service v Chadha, the Court refused to countenance a congressional veto on the Attorney General's decision of whether or not to deport an illegal alien from American soil.[96] Similarly, in Clinton v City of New York, the Court stepped in to preclude

[91] William B Gwyn, 'The Indeterminacy of the Separation of Powers in the Age of the Framers' (1989) 30(2) William and Mary Law Review 263, 263 <http://scholarship.law.wm.edu/wmlr/vol30/iss2/4> accessed 10 January 2018.

[92] *Public Citizen v Department of Justice* 491 US 440 (1989) [485].

[93] *Misretta v United States* 488 US 361 (1989) [382].

[94] ibid.

[95] Tribe, *American Constitutional Law* (3rd edn) (n 78) 139–140.

[96] *Immigration and Naturalization Service v Chadha* 462 US 919 (1983).

executive encroachment on core legislative competencies, despite the legislature's complicity in the encroachment. The Court in that case invalidated legislation, which had been duly passed by Congress, allowing the President to apply "line-item cancellations" to certain provisions of appropriations bills. Effectively, the legislature would have thereby conferred on the executive the ability to carry out unilateral statutory amendments.[97] The Court refused to permit such an augmentation of executive power.

A malleable view of the separation of powers also characterises the approach taken towards relations between the judiciary and the other branches of government. The same article of the Constitution that stipulates the creation of the Supreme Court also vests Congress with the power of establishing such other federal courts as it sees fit.[98] Furthermore, the jurisdiction of the Supreme Court and Federal Courts may be circumscribed "under such regulations as the Congress shall make."[99] Although the Constitution permits these cross-branch interferences, courts will invoke the separation of powers when they perceive intrusion into the judiciary's core jurisdiction beyond what is sanctioned by the Constitution. Thus, in the historic 1872 case of United States v Klein, the Court invalidated legislation directing the Court to interpret a previous law in a manner that would preclude former Confederate soldiers the benefit of compensation for property loss. The Court, in perceiving that it was being "forbidden to give effect to evidence which, in its own judgment such evidence should have", found that Congress had "passed the limit which separates the legislative from the judicial power".[100]

America's adherence to the separation of powers theory is thus a matter of degree. The Federal Constitution faithfully reflects both the first and third principles of the pure theory: the government is divided into executive, legislative, and judicial branches, and the membership of each is kept strictly separate. Indeed, it has been said that the rigidity with which the latter precept has been observed is "the most significant aspect of the doctrine in forming the special character of American government."[101] However, the American Constitution clearly departed from the second principle's strict division of governmental powers along functional lines. As Justice Blackmun observed, writing for the majority in Misretta v United States:

> In adopting this flexible understanding of separation of powers, we simply have recognised Madison's teaching that the greatest security against tyranny… lies not in a hermetic division between the

[97] *Clinton v City of New York* 524 US 417 (1998).
[98] Tribe, *American Constitutional Law* (n 72) 33.
[99] ibid 33.
[100] ibid 39.
[101] ibid 134.

> Branches, but in a carefully crafted system of checked and balanced power within each Branch.[102]

As we have seen, the dilemma of indeterminate categorisation of state powers also necessitated, in the words of the Supreme Court, a "pragmatic, flexible view of differentiated governmental power".[103]

V. The Separation of Powers in Canada

A. Fusion of the executive and legislature

The immediate difference between the American and Anglo-Canadian systems, as they relate to the separation of powers, is the latter's flagrant breach of the pure theory's third principle: the prohibition on the plurality of office. In the Westminster system, the executive Cabinet consists entirely of members of the legislature. This duality was a main criticism of British government during and after the American Revolution. In The Rights of Man, Thomas Paine denounced a system that allowed the same officials to justify in one capacity the measures that they advise and carry out in another; a system in which "the advisers, the actors, the approvers, the justifiers, the persons responsible and the persons not responsible, are the same persons."[104] As for the room such an arrangement leaves for the separation of powers theory, the moguls of British Constitutional theory were convinced the theory had no place whatsoever in the parliamentary system. Walter Bagehot, in The English Constitution, dubbed "erroneous" those descriptions of the Constitution in which "the legislative, executive, and judicial powers are quite divided [such that] each is entrusted to a separate person or set of persons [and] no one of these can at all interfere with the work of the other".[105] In The Law of the Constitution, Dicey also declared the separation of powers an idea "alien to the conceptions of modern Englishmen".[106] Bagehot explained that the concept of a separation of powers, where "ultimate power is different upon different point—now resid[ing] in one part of the constitution, and now in another" is inconsistent with the English system of parliamentary supremacy, in which "the supreme determining power is upon all points the same".[107] Within

[102] *Misretta* (n 93) [381].
[103] *Buckley v Valeo* 424 US 1 (1976) [122], cited in Misretta (n 93) [381].
[104] Thomas Paine, *Rights of Man* (University College Cork 2014) 137 <https://www.ucc.ie/archive/hdsp/Paine_Rights_of_Man.pdf.> accessed 5 November 2017.
[105] Walter Bagehot, *The English Constitution* (OUP 1961) 2.
[106] Dicey (n 47) 337.
[107] Bagehot (n 105) 194–195.

the English system, "[n]o matter whether the question upon which it decides be administrative or legislative, [Parliament] can despotically and finally resolve".[108]

Rather than a separation between legislative and executive powers, Bagehot asserted their fusion to be "[t]he efficient secret of the English Constitution".[109] The connecting link, divined Bagehot, in what is undoubtedly the most famous sartorial metaphor in constitutional law, is the Cabinet—the "buckle which fastens" the legislature to the executive.[110]

From this commentary on the constitution of the United Kingdom, to which the Canadian Constitution is "similar in Principle",[111] it is readily apparent that "the separation of powers in Anglo-Canadian constitutional law is neither explicit nor complete."[112] Peter Hogg, a leading authority on Canadian constitutional law, explains that this stems from the parliamentary system being a form of "responsible government".[113] This appellation denotes the accountability of the executive to the legislative assembly. The Prime Minister holds the premiership by virtue of being the leader of the party commanding a majority in the House of Commons. If the House of Commons passes a motion of no confidence, or if the government is defeated on a vote of sufficient import, the Premier is deemed to have lost the confidence of the majority of the House and cannot continue in office.[114] The Premier must then resign or call elections. This differs from the American system in which the President serves out his term in office regardless of the support of Congress, which is not uncommonly controlled by a party different than the one to which the President belongs.[115]

Thus, the national legislature gives the Cabinet the power to rule and is simultaneously ruled by it, as the Cabinet comprises the leaders of the party predominating in the legislature.[116] Naturally, this marriage of executive and legislative branches is a recurring theme in Canadian jurisprudence. In Attorney General of Québec v Blaikie et al, the Supreme Court ruled that Section 133 of the British North America Act 1867,[117] stipulating French and English language requirements for legislation, also applied to executive Orders in Council issued by provincial governments, as well as to regulations and orders emanating from

[108] ibid 201.
[109] ibid 9.
[110] ibid 11.
[111] Constitution Act 1867 (Canada), 30 & 31 Vict c 3, preamble.
[112] RG Brian Dickson, *The Rule of Law: Judicial Independence and the Separation of Powers: An Address by The Right Honourable Brian Dickson, P.C. Supreme Court of Canada to The Canadian Bar Association* (sn 1985) 6.
[113] Peter W Hogg, *Constitutional Law of Canada*, 2016 Student Edition (Thomas Reuters 2016) ch 9-2.
[114] ibid ch 9-22.2–9-22.3.
[115] ibid 9-3.
[116] ibid 9-22.1.
[117] British North America Act 1867 (Canada), 30 & 31 Vict c 3.

subordinate statutory bodies.[118] In determining that such delegated legislation should be treated the same as legislative enactments for the purposes of Section 133, the Court considered the practical implications of executive-legislative fusion:

> [I]t is the Government which, through its majority, does in practice control the operations of the elected branch of the Legislature on a day to day basis, allocates time, gives priority to its own measures and in most cases decides whether or not the legislative power is to be delegated and, if so, whether it is to hold it itself or to have it entrusted to some other body.[119]

Accordingly, the Court found that legislative powers delegated by the legislature to the executive, "which is part of itself", must be viewed "as an extension of the legislative power of the legislature" such that "the enactments of the Government under such delegation must clearly be considered as the enactments of the Legislature".[120]

Even where asserting a distinction between executive and legislative branches might appear advantageous, their symbiosis in the parliamentary system means that the executive often cannot escape being implicated in actions of the legislature. For this reason, in Wells v Newfoundland the Supreme Court held that the Newfoundland Government could not point to the passage of a provincial statute as an event that had frustrated its contract with the Plaintiff so as to relieve it from liability. The Court refused to entertain such a claim from the Province, given that "the same individuals control both the executive and legislative branches of government… therefore it is disingenuous for the executive to assert that the legislative enactment of its own agenda constituted a frustrating act beyond its control".[121]

A more pernicious side of the parliamentary arrangement came to into focus in Canada (Auditor General) v Canada (Minister of Energy, Mines and Resources).[122] In that case, Petro-Canada, a Crown Corporation, had refused to furnish information to the Auditor General who was investigating on Parliament's behalf the propriety of a major asset purchase made with appropriated funds. The Governor in Council, which approved Petro-Canada's annual budget, refused to order Petro-Canada's compliance with the Auditor General's request for disclosure.

[118] *Attorney General of Québec v Blaikie et al* [1981] 1 SCR 312 [333].
[119] ibid [320].
[120] ibid.
[121] *Wells v Newfoundland* [1999] 3 SCR 199 [53].
[122] *Canada (Auditor General) v Canada (Minister of Energy, Mines and Resources)* [1989] 2 SCR 49.

In these circumstances, the Auditor General's only statutory recourse, set out in Section 7(1)(b) of the Auditor General Act 1976-77,[123] was to report to Parliament. That body, however, he found unwilling to take the Government to task for its lack of transparency. The Court, to which the Auditor General turned for redress in light of Parliament's lethargy, refused to provide an alternative remedy to the parliamentary reporting procedure. The fact that the thoroughly conservative Parliament was indisposed to scrutinise questionable expenditures made by the Mulroney Administration was merely a symptom of the Westminster system, and, as such, not reviewable by the Court. The Court could only shrug its shoulders at the realities of the system and inform the Auditor General that he must do the same:

> It is of no avail to point to the fusion of powers which characterizes the Westminster system of government. That the executive through its control of a House of Commons majority may in practice dictate the position the House of Commons takes on the scope of Parliament's auditing function is not… constitutionally cognizable by the judiciary.[124]

As the Supreme Court made clear, for better or worse, there is no separating the executive from the legislature in Canada's parliamentary system.

Like Bagehot and Dicey before him, Peter Hogg has observed the fate to which the separation of powers theory is condemned in a parliamentary system: "The close link between the executive and legislative branches which is entailed by the British system is utterly inconsistent with any separation of the executive and legislative functions."[125] The Supreme Court of Canada has often agreed with this view. Indeed, on multiple occasions[126] the Court has cited the following passage from Hogg:

> There is no general 'separation of powers' in the Constitution Act, 1867. The Act does not separate the legislative, executive, and judicial functions and insist that each branch of government exercise only 'its own' function. As between the legislative and executive branches,

[123] Auditor General Act 1976–77 (Canada).

[124] ibid [103].

[125] Hogg (n 113) ch 14-5.

[126] *Douglas/Kwantlen Faculty Assn v Douglas College* [1990] 3 SCR 570 [601] (La Forest J); *MacMillan Bloedel Ltd v Simpson* [1995] 4 SCR 725 [52] (McLachlin J, dissenting); *Re Residential Tenancies Act*, 1979 [1981] 1 SCR 714 [728] (Dickson J). See also *Reference re Secession of Quebec* [1998] 2 SCR 217 [15].

> any separation of powers would make little sense in a system of responsible government; and it is clearly established that the Act does not call for any such separation. As between the judiciary and the two political branches, there is likewise no general separation of powers. Either the Parliament or the legislatures may by appropriate legislation confer non-judicial functions on the courts and… may confer judicial functions on bodies that are not courts.[127]

Incredibly, notwithstanding the above, the separation of powers theory is not uncommonly referred to in Canadian jurisprudence as a "fundamental principle of the Canadian Constitution."[128] Still more ironic, given the pronouncements of Bagehot and Dicey on the subject, is where the Supreme Court has located the source of this principle of Canadian law. "The separation of powers", wrote Justice McLachlin (as she then was) in her dissenting opinion in Cooper v Canada, "was incorporated into the Canadian Constitution by the Constitution Act 1867, through that provision's reference to a constitution 'similar in Principle to that of the United Kingdom.'"[129] The majority in Harvey v New Brunswick pointed to the same preambular clause as the source of the separation of powers in the Canadian Constitution, declaring the principle to be "inherent in British parliamentary democracy".[130]

B. Separation of the judiciary from the political branches

The separation of powers, to the extent that it applies in Canada, concerns the relationship between the courts and the political branches of government. Though a separation of powers in toto is incompatible with Canada's state structure, there is a conviction that some measure of separation is fundamental to the state's institutional arrangement. Justice Dickson, formerly of the Supreme Court, observed that, "some of the powers in the constitution were and still are, so separated that their holders have autonomous powers."[131] Justice Dickson explained that "judges have power of this nature because, being entrusted with the maintenance of the supremacy of the law, they are and always have been regarded

[127] Hogg (n 113) 7–37.
[128] *Provincial Judges Reference* (n 37) [138]. See also Cooper (n 37) [28] and *Doucet–Boudreau v Nova Scotia* 2003 SCC 62 [94].
[129] *Cooper* (n 37) [22]. See also *Canada v Vaid* 2005 SCC 30 [21]; Dickson (n 112) 6.
[130] *Harvey v New Brunswick* [1996] 2 SCR 876 [68].
[131] Dickson (n 112) 6.

as a separate and independent part of the constitution".[132] As alluded to by Justice Dickson, the application of the separation of powers theory in Canada has been largely, if not exclusively, focused on "the relationships between the legislature and the executive on the one hand, and the judiciary on the other".[133] The dissent in Doucet-Boudreau v Nova Scotia summed up the extent of the theory's application in Canada accordingly: "Our Court has strongly emphasised and vigorously applied the principle of separation of powers in order to uphold the independence of the judiciary" and ensure that courts "as a general rule, avoid interfering in the management of public administration."[134]

However, not even in this respect does the principle hold fast in Canadian practice. One illustration of this is the Supreme Court's power to render advisory opinions. This is a traditionally "executive" function, performed by the Attorney General and other law officers of the government.[135] However, the Court has held that, because "the Canadian Constitution does not insist on a strict separation of powers[,] Parliament and the provincial legislatures may properly confer other legal functions on the courts, and may confer certain judicial functions on bodies that are not courts."[136] Though the function may be "legal", the Court recognised it is not "judicial", being outside the framework of adversarial litigation or genuine controversy.[137] As the Court found nothing objectionable in this, clearly, as in the United States, the second principle of the pure separation of powers theory, mandating a strict separation of powers along functional lines, is not followed in Canada, even as between the judiciary and the political branches.

C. Normative value: an unhelpful guide to "proper spheres"

The object of the separation of powers theory in Canadian constitutional theory is to maintain some manner of separation between the courts and political branches. However, the theory provides little guidance as to how each side's proper sphere is to be delineated. Broad statements invoking the theory, but offering little utility in actually delimiting institutional boundaries, are found in a variety of Supreme Court decisions.[138] In the Provincial Judges Reference, the Court held that the separation of powers requires "the preservation of the basic structure" of

[132] ibid.
[133] *Provincial Judges Reference* (n 37) [140]. See also Doucet–Boudreau (n 128) [108].
[134] *Doucet–Boudreau* (n 128) [109]–[110].
[135] Hogg (n 113) 8–19, 7–37.
[136] *Secession Reference* (n 126) [14].
[137] ibid [15].
[138] *New Brunswick Broadcasting Co v Nova Scotia* [1993] 1 SCR 319 [389].

government.[139] In Canada v Vaid, the Court stated that, by virtue of the separation of powers principle, "[e]ach of the branches of the state is vouchsafed a measure of autonomy from the others."[140] In Cooper v Canada, the Court declared that "the separation of powers requires that certain functions be exclusively exercised".[141] But what functions? It is all very well to state that "the judiciary must be free from encroachment by government upon matters within its proper sphere" and, equally, "the judiciary must not encroach upon the proper domain and jurisdiction of government".[142] But with no functional guideline, the "proper domain" of each branch is indeterminate.

The divergence of opinion between the majority and dissent in Doucet-Boudreau, to take one prominent example, demonstrates how uncertainty in demarcating the legitimate sphere of each branch is in no way attenuated by appealing to the separation of powers theory. Doucet-Boudreau concerned a decision by the Nova Scotia Supreme Court to implement a novel remedy in seeking to uphold the French-speaking minority's language rights guaranteed by the Canadian Charter of Rights and Freedoms.[143] Francophones had been promised French language schools that, for years, failed to materialise, while their language increasingly faced the dangers posed by assimilation. To protect the Charter rights at issue in the face of the provincial government's inaction, the trial judge set compliance deadlines for the government and retained personal jurisdiction to order hearings to verify progress was being made.[144]

The majority upheld the lower court's remedy despite its dubiously executive character. The majority began by paying lip service to the existence of a separation of powers, admonishing that "courts must be sensitive to their role as judicial arbiters and not fashion remedies which usurp the role of the other branches of governance".[145] Then, as if to illuminate a threshold on which it really cast little, if any, light, the majority decreed that, "[d]eference [to the other branches] ends, however, where the constitutional rights that the courts are charged with protecting

[139] *Provincial Judges Reference* (n 37) [108].
[140] *Vaid* (n 129) [21].
[141] *Cooper* (n 37) [13]. See also *Provincial Judges Reference* (n 37) [139].
[142] Dickson (n 112) 11–12.
[143] Canada Act 1982, Schedule B Constitution Act 1982 (Canada), Part I Canadian Charter of Rights and Freedoms.
[144] *Doucet–Boudreau* (n 128) [5]–[15].
[145] ibid [34].

begin".[146] The majority went on to confirm the separation of powers principle as the supreme tool of pragmatism in Canadian jurisprudence:

> A remedy may be appropriate and just notwithstanding that it might touch on functions that are principally assigned to the executive. The essential point is that the courts must not, in making orders under s 24(1), depart unduly or unnecessarily from their role of adjudicating disputes and granting remedies that address the matter of those disputes.[147] [emphasis in italics added]

Thus, even the vague standard of respecting each branch's proper sphere is subject to the broad qualification that judicial transgressions are permissible where necessary to uphold the Charter.

The dissent took a harder line on the requirements of the separation of powers. The dissent was of the mind that "[d]espite—or, perhaps, because of—the critical importance of their functions, courts should be wary of going beyond the proper scope of the role assigned to them in the public law of Canada".[148] The dissent concluded that, "[b]y purporting to be able to make subsequent orders [after disposing of the matter before him], the trial judge would have assumed a supervisory role which included administrative functions that properly lie in the sphere of the executive".[149]

The two opinions in Doucet-Boudreau are diametrically opposed. Significantly, the Court as a whole approached the issue with a similar understanding of the separation of powers and its elasticity. The majority recognised that "extend[ing] the court's jurisdiction beyond its proper role… will breach the separation of powers principle",[150] while the dissent agreed that the separation of powers "flexibly delineates the domain of court action".[151] However, the fact remained that the separation of powers theory offered no guidance on the court's proper role or domain. The dissent therefore viewed the lower court's presumption to supervise executive compliance with its decision as an encroachment on executive power and a violation of the separation of powers. Conversely, the majority found that such an injunction was central to the court's ability to fashion Charter remedies, similar to contempt proceedings, garnishments, and writs of seizure.[152] What the majority perceived as the exercise of a core judicial right, the dissent viewed as an

[146] ibid [36].
[147] ibid [56] (emphasis added).
[148] ibid [106].
[149] ibid.
[150] ibid [105].
[151] ibid [94] (emphasis added).
[152] ibid [70].

unacceptable intrusion into the executive's core competencies. The divergence in the Court's opinion is symptomatic of the lack of guidance proffered by the theory.

In Canada, the "separation of powers" is not used to indicate a division of governmental power according to function, nor a prohibition on the plurality of office, nor the institutional restraints on each branch that such an arrangement is said to give rise to. It is of little to no practical use in defining the appropriate boundaries between the courts and the other branches or identifying the powers that might lie within those boundaries. In short, the separation of powers is merely a catch-all phrase to refer to Canada's existing arrangement of state powers. It seems that, as Sir William Ivor Jennings, one of the twentieth century's leading authorities on constitutionalism, declared,

> [I]f a political principle which has some basis in reason receives general acceptance and can be formulated in a neat phrase, it becomes a reason in itself; its original justification is forgotten, and it is used for purposes for which it was never intended.[153]

D. Saving grace: the cautionary function

Despite the criticism that, in Canada, the separation of powers principle lacks normative value for institutional arrangement, the existence of the principle does have one redeeming quality. The persistence of the idea of a "separation of powers" in the ethos of Canadian constitutionalism encourages a measure of caution before either the courts or the political branches of government decide to undertake a course of action with the potential to affect the existing constitutional structure. The unarticulated "proper role" of each branch, whatever it may be, is duly considered before any such action is taken. In Vaid, for instance, the idea that a separation of powers must be maintained imbued the Court with a sense of prudence before rendering a decision with the potential to interfere with parliamentary privilege and hence Parliament's bona fide functions. At the same time, the principle was flexible enough that the Court was not compelled to countenance such an injustice as Parliament "deny[ing] its employees human rights protections which Parliament itself imposed on every other federal employer".[154] The point, however, is the value of the initially cautious posture assumed by the Court. If the Court was not conditioned to observe such caution, the legislature might be deprived of control over its own procedure. In such an eventuality, "inefficiency would result from

[153] Jennings (n 42) 25.
[154] *Vaid* (n 126) [2].

the delay and uncertainty would inevitably accompany external intervention".[155] Herein lies the value of the separation of powers principle in Canada, as the Court's caution proceeds partly from a notion that there is a separation of powers among the branches of the Canadian government that must be respected. The separation of powers is not a matter of ideological orthodoxy, nor, as the Court has stated, is it a sign of respect for parliamentarians.[156] It is a principle that militates against the weakening of any branch of government, not so much to prevent the strengthening of another branch, but to ensure the continuing effectiveness and efficiency of all branches.

VI. Comparison: Canada and the United States

One key difference between Canadian and American practice concerning the separation of powers is that, in Canada, there is less emphasis on allocating state powers on the basis of function. This method of allocation is, in American constitutional theory and practice, partially maintained outside the established regime of checks and balances. Thus, references to state actions that are, for instance, "quintessentially executive",[157] are not uncommon in American jurisprudence. However, this type of labelling, as a precursor to appropriate allocation, is not reflected in Canadian jurisprudence, which evinces greater concern for branches carrying out their "proper roles" than respecting functional divisions. In executing its perceived proper role, one branch may legitimately infringe on the functions of another. This variance in state practice is evidenced by the Supreme Court of Canada's power to render advisory opinions—a prerogative not shared by its American counterpart because the function is not, strictly speaking, judicial.[158]

Apart from this difference, the principle serves a relatively similar purpose in both countries: preventing an undue encroachment of one branch upon another, such that one branch is unacceptably weakened, or another unacceptably strengthened, producing either the threat of tyranny or inefficiency.

The core comparison therefore comes down to the major difference between the two systems of government as they relate to the separation of powers theory. This difference centres on Canada's patent violation of the pure theory's third principle prohibiting plurality of office, a principle so vigilantly adhered to in the United States. The parliamentary arrangement, according to Bagehot, wards off political deadlock—that pitfall of the American system that has led some constitutional scholars to declare, "the fundamental problem, in trying to make the

[155] ibid [29].
[156] ibid.
[157] *Morrison* (n 80) [705]–[706].
[158] *Secession Reference* (n 126) [13]–[15].

government of the United States work effectively, is not to preserve the separation of powers, but to overcome it."[159] A statutory initiative emanating from the President may be so altered by demands of individual congressmen with their own distinct interests that, if the bill does emerge from the congressional committees, the series of compromises it has been made to reflect often leave it bearing no resemblance to what was initially proposed.[160] Furthermore, while Congress presides over the law, the President carries out its administration, and "the President does not have to exert himself in carrying out laws he doesn't approve of".[161] Though the President may be impeached for gross negligence, "between criminal nonfeasance and zealous activity there are infinite degrees".[162] In the extreme, such a dynamic may devolve into what Professor Bruce Ackerman of Yale Law School terms a "crisis in governability", characterised by "endless backbiting, mutual recriminations, and partisan deadlock" with each branch of government using its constitutional prerogatives to frustrate the activities of the others.[163]

In Canada, not only does the separation of powers principle not prevent the subjugation of one branch by another as it does in America, the principle actually protects such subjugation. The Supreme Court of Canada in the Provincial Judges Reference explained that one aspect of the Canadian separation of powers is the protection of the "hierarchical relationship between the executive and the legislature, whereby… once legislatures have made political decisions and embodied those decisions in law, it is the constitutional duty of the executive to implement those choices".[164] In Canada, the separation of powers thus allows for overlap of the executive and legislative branches and also maintains the subjugation of the former by the latter. This subjugation guards against the kind of foot-dragging and mutual-frustration techniques that are facilitated by the executive-legislative divide in the American version of the separation of powers. Government policy, perceived Bagehot, "acts by laws—by administrators; it requires now one, now the

[159] Goldwin and Kaufman (n 89) 138–139.
[160] Ackerman (n 8) 254.
[161] Bagehot (n 105) 197.
[162] ibid.
[163] Norman Dorsen, Comparative Constitutionalism: Cases and Materials (Thomson/West 2010) 215.
[164] *Provincial Judges Reference* (n 37) [139].

other" and the excellence of the British Constitution is that it has achieved unity between the two and "in it, the sovereign power is single, possible, and good".[165]

VII. Conclusion

The meaning of the "separation of powers" has altered significantly since the theory's inception during the English Civil War and early experiments with the theory in the newly liberated American colonies. The deficiencies of the pure version of the theory quickly manifested themselves, and the mixed-government mechanism of checks and balances was superimposed onto the theory to compensate. Despite contrary misconceptions, a pragmatic and flexible approach rather than a strict approach toward the separation of powers was intended by the Framers of the American Constitution. Generally, such an approach has also prevailed in American jurisprudence. While there is still impetus to allocate state powers to each branch of government according to function, a fundamental element of the American Constitution—its checks and balances—precludes any such rule being strictly adhered to. The indeterminate nature of many state activities, which defy the requisite labelling, also precludes such a rule from being faithfully observed. The United States government is, however, divided into executive, legislative, and judicial branches, and rigidly enforces a ban on concurrent membership in any two branches. This is more than can be said for Canada, as far as observance of the pure theory of the separation of powers is concerned.

The Anglo-Canadian system of parliamentary government is fundamentally at odds with the prohibition on the plurality of office, given that the executive is composed exclusively of members of the legislative assembly. In Canada, the theory is interpreted exclusively in terms of the relationship between the courts on the one hand, and the two political branches of government on the other. The separation of powers does not refer to anything more concrete than the institutional arrangements that characterise the Canadian government at any given time. The doctrine serves as a cautionary brake upon any branch contemplating any sort of foray outside its conventional sphere of activity. While this strays far indeed from the separation of powers as originally conceived, the theory, so understood, serves no small benefit in Canadian constitutional practice.

It is difficult to directly compare the separation of powers in Canada and the United States; to do so is to deal in two different currencies. In both nations, the principle has the same salutary effect of cautioning against adventurist institutional action. However, in the United States, the separation of powers has been interpreted to require a rigid separation of the political branches from one another, and the simultaneous arming of each branch with the constitutional means of foiling the

[165] Bagehot (n 105) 202–203.

others' designs. While this approach to the separation of powers may prevent an undesirable accumulation of power in any single branch of government, it also has the unwelcome effect of facilitating political stalemate unless the branches exhibit a sufficient degree of cooperation and willingness to compromise. The Canadian approach may have the advantage of avoiding this predicament.

Caution—"Do Not Cross": Drawing a Perimeter on Police Deception

Alexis Wilt*

I. Introduction

Imagine you have just become the victim of a second burglary of your home. This time, the burglar tripped your newly installed alarm system and police respond to the scene. After you have spoken with the police about the incident, they leave. Later, two uniformed officers show up on your doorstep claiming they are there to follow up on the investigation into the burglary. In the hope that they will catch the criminal and recover your belongings, you allow them inside your home. Believing that they are there to help you, you allow them access into private areas of your home while they dust multiple places for fingerprints and probe further and further into your house.

Now imagine a similar situation, but instead, one of the two officers identifies himself as a member of the Fraud Task Force there to investigate you for fraud, as well as for the burglary, and the other officer introduces himself as an agent of the Secret Service, also there to investigate possible fraud. Additionally, imagine the officers have informed you that they have apprehended the suspect for the burglary you reported and that he has confessed to the crime. Evidently, these two scenarios present two very different situations. Provided with this additional information, would you still consent to the officers entering your home to conduct a search?

The first scenario is what took place in the case of *United States v Spivey* (*Spivey*).[1] Police officers devised an extravagant ruse designed to induce the defendants to consent to a search of their home. Chenequa Austin led officers on an in-depth search of her house, believing that the officers were there to investigate a burglary of her home and aid her in this time of vulnerability. Instead, it turned out that the

* Law Student, Florida State University College of Law. B.A. in Criminology and Sociology (Florida State University). I would like to acknowledge Professor Wayne Logan, my mentor throughout the writing process of this article and a source of inspiration to all criminal law students, Chris Villa for his great feedback and critiques throughout the drafting process, and my family for their constant motivation and support.

1 *United States v Spivey* 861 Federal Reporter 3d 1207 (Eleventh Circuit 2017).

officers were chiefly there to investigate Austin and Spivey for possible credit card fraud which the suspect of the burglary, whom they had already caught, had put them on notice of. When officers informed Austin of their real reason for searching the residence, her cooperation in the investigation immediately ceased, as she realised the officers were not there to aid her as a *victim* of a crime but to investigate her as a *suspect* of a crime. Should courts allow this kind of police deception? If so, where do courts draw the line on how much deception is permissible?

Part I of this article explores the constitutional protections provided by the Fourth Amendment and some of the exceptions the courts have carved out over the years. One particularly important exception, and the forefront issue in *Spivey*, is the 'consent exception' to the warrant requirement. This Part will delve into the dynamics of the consent exception and how the courts evaluate the voluntariness of a subject's consent through consideration of the totality of the circumstances. There are multiple factors for the court to consider when determining the voluntariness of consent, such as the use of coercion to induce consent.

Part II lays out the procedural posture of the central case discussed in this article—*Spivey*. It explores the District Court's findings that Austin's consent was voluntary and that any potential problem with the defendant's initial consent was later cured by Spivey's subsequent signing of a written waiver of a search warrant. Next, it examines the Court of Appeals' majority decision to affirm the lower court's finding that Austin's consent was voluntary, as well as the dissenting opinion.

Part III provides an in-depth interpretation of current case law pertaining to the issues at hand in *Spivey* and an insight into how existing case law should be applied to the facts of this case. First, it will explore what constitutes coercion and how its presence may render consent involuntary. The use of blatant misrepresentations may deprive an individual of their ability to accurately assess a situation which, in turn, renders any subsequent consent involuntary, given that the subject is unable to knowingly and voluntarily give consent. Next, it discusses how consent cannot later be 'cured' as the District Court suggested in its ruling. Written consent given after the search has already been conducted is analogous to submission to authority rather than voluntary consent. Lastly, if any evidence is produced through an unlawful, warrantless search, it is considered 'fruit of the poisonous tree' and may be suppressed from being entered into evidence at trial.

Part IV discusses the potential consequences of the precedent set forth in *Spivey*. By upholding this decision, courts are permitting officers to deliberately circumvent the warrant requirement provided in the Fourth Amendment in a manner that undermines the public's trust in law enforcement. Important practical and doctrinal ramifications may result from such a decision. Although police deception is common practice in criminal investigations, "[t]hey raise deep social and ethical problems that provoke concern about the acceptable role of police

behaviour within the parameters of the Fourth Amendment".[2] This Part highlights some of the important policy concerns courts should consider when determining the parameters of acceptable police deception—concerns such as citizens' trust of law enforcement officers and the condoning of police stratagems that contradict the protections intended by the Framers of the Constitution.

Part V suggests three rules the courts should adopt to clarify some of the grey areas of police use of deception to gain consent. These rules are expected to create a more structured analysis on which courts shall base their decisions. The first proposed rule envisages that officers should not be allowed to impersonate, or act under the guise of, another government official in an effort to induce consent. Such a façade is analogous to a fraudulent claim of authority, and prior case law clearly discourages the use of deception by government agents while acting under disclosed official capacity. The second proposed rule requires uniformed officers to disclose all investigations they intend to conduct prior to obtaining consent from a subject. This rule promotes a 'meeting of the minds' regarding the kind of search the suspects are consenting to, and fosters the public's trust in officers by discouraging officers from deceiving subjects into believing they are there to aid them when in reality they are there to investigate the subjects of the search. This rule would require courts to evaluate a number of factors in determining whether the government had intentions of conducting an investigation which had not been disclosed to the subject prior to obtaining their consent. The third proposed rule is offered as an alternative to the second one. It suggests requiring officers to disclose their specific assignments and tasks forces within the agency for which they are acting at the time of the search. Requiring such disclosure prior to obtaining consent promotes transparency between government officials and citizens, while eliminating some of the potential for officers to deceive subjects into consenting.

II. The Consent Exception to the Fourth Amendment

The Fourth Amendment of the Constitution provides "[t]he right of the people to be secure in their persons, *houses*, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause".[3] Courts have emphasised the importance of protecting the heightened expectation of privacy that citizens have in their homes.[4] The Fourth Amendment generally presumes that warrantless searches of a person's home

[2] Elizabeth N Jones, 'Professional Article: The Good and (Breaking) Bad of Deceptive Police Practices' (2015) 45 New Mexico L Rev 523, 523.

[3] US Constitution Amendment 4 (emphasis added).

[4] *Welsh v Wisconsin* 466 US 740 (WI 1984) 748 ("It is axiomatic that the physical entry of the home is the chief evil against which the wording of the Fourth Amendment is directed.").

are unlawful.[5] Courts, however, have carved out some exceptions to the warrant requirement employed by the Fourth Amendment. One well recognised exception to the warrant requirement is when a person voluntarily gives consent for an officer to conduct a search—the consent exception.[6]

A lynchpin of the consent exception is that the subject's consent must be given *voluntarily*. The government bears the burden of proving this by a preponderance of the evidence.[7] For consent to be voluntary, it must first be "the product of an 'essentially free and unconstrained choice'".[8] Secondly, courts must evaluate each case individually based on the totality of the circumstances.[9] In evaluating the totality of the circumstances, there is no single factor that controls the analysis.[10] Instead, there are a number of factors the court must consider.[11] The Court in *Purcell* enumerated some of these factors: (a) coercive police procedures; (b) the defendant's cooperation with officers; (c) the defendant's understanding of his right to refuse; (d) the defendant's education level and intelligence; and (e) the defendant's belief that officers will not find incriminating evidence.[12]

The Court in *Schneckloth v Bustamonte* determined that consent is not voluntary when it is "the result of duress or coercion, express or implied".[13] The term 'coercion' is not restricted to physical coercive measures. In the case of *Blackburn v Alabama*, the Court recognised that coercion may be both physical and mental, and that certain refined methods of persuasion may be tantamount to coercion.[14] In other words, depending on the surrounding circumstances, consent given through sophisticated means of persuasion—such as deception—may render a person's consent involuntary.

One facet of voluntary consent that remains uncertain is the degree of deception that law enforcement is permitted to use to induce consent. Deception is not only commonplace in law enforcement, but is also encouraged in law

5 *Payton v New York*, 445 US 573 (NY 1980) 586 ("[S]earches and seizures inside a home without a warrant are presumptively unreasonable."). See also *Kyllo v US*, 553 US 27 (2001) 31 ("We know that with few exceptions, the question whether a warrantless search of a home is reasonable and hence constitutional must be answered no.").

6 *United States v Garcia* 890 Federal Reporter 2d 355 (Eleventh Circuit 1989) 360. See also *Illinois v Rodriguez* 497 US 177 (1990) 181.

7 *US v Yeary* 740 Federal Reporter 3d 569 (Eleventh Circuit 2014) 581.

8 *US v Purcell* 236 Federal Reporter 3d 1274 (Eleventh Circuit 2001) 1281 (citing *Schneckloth v Bustamonte* 412 US 218 (1973) 225).

9 *Yeary* (n 7) 581.

10 *Schneckloth* (n 8) 226.

11 *Purcell* (n 8) 1281.

12 ibid.

13 *Schneckloth* (n 8) 227, 248.

14 *Blackburn v Alabama* 361 US 199 (1960) 206.

enforcement training manuals,[15] and is routinely accepted by the courts.[16] In fact, the United States Supreme Court, when presented with the opportunity to craft a bright-line rule on the use of police deception to induce consent, refused to do so in fear that it would hamper the government's ability to carry out criminal investigations.[17] Courts must, however, evaluate each case on its own merits to determine whether the police deception involved is extreme enough to render the individual's consent involuntary under the Fourteenth Amendment's Due Process Clause.[18]

Often discussed in relation to the voluntary consent exception is the 'plain view' doctrine. In *Horton v California*, the court outlined a three-part analysis which police officers must satisfy before a seizure is deemed constitutional under the plain view doctrine.[19] First, an officer must be lawfully present at the place where the evidence was observed in plain sight.[20] Secondly, the officer must have a lawful right of access to the evidence.[21] In other words, the officer must not be required to do any further intrusion to retrieve the evidence. Lastly, the incriminating character of the evidence must be "immediately apparent".[22]

The first prong of the analysis is important for cases involving the issue of police deception because, if the initial consent to a search is rendered involuntary due to the degree of deception, then any evidence procured by the search is considered 'fruit of the poisonous tree'. In 1914, the Court in *Weeks v United States* held that evidence obtained in violation of the Fourth Amendment must be excluded from federal courts.[23] The dual rationale behind this exclusionary rule was to deter police misconduct and preserve judicial integrity which required that courts do not sanction these illegal searches by admitting the fruits of illegality into evidence.[24] Nonetheless, one issue with this decision was that it applied only to federal cases. It was not until 1961, in *Mapp v Ohio*, when the Court finally held

[15] Fred E Inbau, John E Reid, Joseph P Buckley and Brian C Jayne, *Criminal Interrogation and Confessions* (5th edn, Jones & Bartlett Learning 2011) for descriptions of interrogation techniques taught to officers to use when interrogating suspects.

[16] Jones (n 2) 523.

[17] *Lewis v US* 385 US 206 (1966) 210 (refusing to craft a *per se* rule holding that deception by law enforcement agents is unconstitutional because "[s]uch a rule would, for example, severely hamper the Government in ferreting out those organised criminal activities that are characterized by covert dealings with victims who either cannot or do not protest").

[18] ibid 212 ("[I]n this area, each case must be judged on its own particular facts.").

[19] *Horton v California* 496 US 128 (1990) 136.

[20] ibid ("It is, of course, an essential predicate to any valid warrantless seizure of incriminating evidence that the officer did not violate the Fourth Amendment in arriving at the place from which the evidence could be plainly viewed.").

[21] ibid 137.

[22] ibid 136.

[23] *Weeks v United States* 232 US 383 (1914) 398.

[24] ibid.

that the exclusionary rule applies to the states as well as to the federal jurisdiction.[25] This case set the precedent that an officer—state or federal—making an initial unreasonable intrusion is not in a valid position to make an observation regarding the incriminating evidence and, therefore, the evidence must be suppressed.[26]

III. Laying the Foundation

The case presented in the introduction of this article, *Spivey*,[27] introduces a perfect avenue for courts to unravel this messy and sensitive area of unsettled law—police deception to induce consent. It is important that, before delving into the fundamentals of the case, the following aspects are reviewed in detail: (a) the facts of the case; (b) its procedural history; and (c) the approaches that other courts have taken on this issue.

A. Facts

Austin and Spivey were the victims of two burglaries by Caleb Hunt. During the second burglary, Hunt tripped a newly installed alarm system and the police responded. Austin spoke with the responding officers about the burglar. Police were able to apprehend the suspect, who confessed to the burglaries and further informed the officer that there was considerable evidence of credit card fraud and a lot of high-end merchandise in their house. The police department that had detained Hunt then closed the case.

Upon receiving this tip, a team of ten law enforcement officers methodically devised a plan which would allow them to avoid obtaining a search warrant to search Austin and Spivey's residence.[28] The contrived plan consisted of sending two members of the South Florida Organised Fraud Task Force to the residence[29]—Detective Alex Iwaskewycz and Agent Lanfersiek. The plan was for these two

[25] *Mapp v Ohio* 367 US 643 (1961) 655 ("Since the Fourth Amendment's right of privacy has been declared enforceable against the States through the Due Process Clause of the Fourteenth, it is enforceable against them by the same sanction of exclusion as is used against the Federal Government.").

[26] Thomas K Clancy, *The Fourth Amendment: Its History and Interpretation* (2nd edn, Carolina Academic Press 2014) s 7.4.4.4.1.

[27] *Spivey* (n 1).

[28] ibid 1219 (Agent Lanfersiek testified that rather than getting a warrant, he and about ten other officers gathered for a planning meeting during which they "made a decision to come up with the methodology of employing the ruse".).

[29] ibid 1221 ("The task force pairs Secret Service agents with local detectives to combat financial crimes in the Southern District of Florida.").

fraud-specialised officers to go to the residence of Austin and Spivey under the pretence of following up on the burglary investigation.

Upon arrival at the residence, officers told Austin they were there to follow up on the burglary investigation. The officers testified that Austin was "genuinely excited" and "relieved" to have the police there to follow up on the burglary[30] and consented for them to enter the residence. To keep up the ruse and avoid suspicion of their fraud investigation, Agent Lanfersiek dressed up as a crime-scene technician for the police department, and pretended to dust for fingerprints as Austin led him through different rooms of the home. Well aware that the suspect had been caught and arrested, both officers chose not to disclose this fact to Austin as they continued to request permission to probe further into private areas of the home. In some of these areas, Agent Lanfersiek saw evidence of credit card fraud in plain view.

Officers then decided to separate Austin and Spivey and speak to each of them one-on-one. This is when the façade ended and the officers disclosed the true nature of their investigation. Detective Iwaskewycz asked Austin about the evidence he saw in the bedroom. Austin immediately became reluctant to cooperate with the detective. Detective Iwaskewycz became aware that Austin was not likely to provide the consent for a full search that he was seeking so he phoned a co-worker to run a name check on Austin. The colleague informed him there was an unrelated outstanding warrant on Austin and the detective arrested her.

While Austin was being questioned outside, Spivey remained inside with Agent Lanfersiek. Spivey continued to cooperate with officers and signed two consent forms that granted the officers permission to conduct a full search of their home, computers, and cell phones. The subsequent search revealed high-end merchandise, MDMA, a loaded handgun, an embossing machine, a card reader-writer, and at least seventy-five counterfeit credit cards.

B. Procedural posture

The case first appeared in the United States District Court for the Southern District. After a federal grand jury returned an indictment against Austin and Spivey, the defendants moved to suppress all evidence obtained as a result of the officers' "entry into the Austins' residence… by fraud… which vitiated any consent".[31] The district court denied the defendants' motion to suppress and rejected a "[b]right line rule that any deception or ruse vitiates the voluntariness of a consent to search".[32] In its reasoning, the district court focused on the evidence that Austin wanted to cooperate with the officers in solving the burglaries because

[30] ibid 1219.
[31] ibid 1212.
[32] ibid.

expensive shoes had been stolen.[33] The court also found that any issue with Austin's initial consent was "cured" by Spivey's subsequent signing of a written waiver for a search warrant.[34] The district court further determined the government proved its burden, by clear and positive testimony, that the defendants' consents were "[v]oluntary, unequivocal, specific, intelligently given, and uncontaminated by duress or coercion".[35] Following this denial of the defendants' motion to suppress, Austin and Spivey conditionally pleaded guilty.

On appeal, the majority affirmed the lower court's ruling that the consents were voluntary and the evidence procured as a result of the search was not to be suppressed. The majority opinion seemed to centre around the notion that, by seeking the help of law enforcement, the defendants had voluntarily exposed themselves to the risk that officers would discover evidence of their own illegal activities.[36] According to the majority, a warning of the right to refuse consent is less relevant in this context, as it is easier to refuse consent when police are there to help than when they are there for adversarial reasons.[37] The majority also downplayed the deceptive effects the ruse may have had on the defendants' consent by brushing the ploy off as "perhaps silly".[38]

The dissenting opinion on appeal focused on three main aspects of the case to evaluate the totality of the circumstances, ultimately determining that the deceptive ruse employed by the officers rendered the defendants' consent involuntary. The first feature of the case on which the dissent focuses is the fact that officers only obtained the defendants' consent to enter their home through deliberate misrepresentation of their authority. The second important aspect of the case was that the officers not only engaged in a misleading ruse, but methodically planned it to circumvent the warrant requirement provided in the Fourth Amendment. Finally, the dissent points out Spivey's refusal to cooperate with law enforcement officers upon learning the true nature of their investigation, showing that she

[33] ibid.
[34] ibid 1212.
[35] ibid.
[36] ibid 1216 (stating that voluntary consent can carry with it the risk that officers may discover evidence of criminal behaviour).
[37] ibid.
[38] ibid 1215.

would not have allowed them into her home if she had known the real intentions behind their investigation.

C. Standard of review

If this case reaches review by the Florida Supreme Court, the Court must review the District Court's legal conclusion on voluntariness *de novo*.[39] When a court reviews a case *de novo*, it uses the trial court's record but reviews the questions of law without deference to the lower court's ruling. In other words, on appeal, the Florida Supreme Court would review the finding of voluntariness regarding Spivey's consent without taking the lower court's decision into consideration.

IV. Unraveling Deception with Case Law

As the dissent points out, litigation in this case could have been completely avoided if the officers had simply obtained a warrant to search Spivey's residence. Courts presume that searches and seizures inside a home without a warrant are unreasonable.[40] This presumption of unreasonableness is also why the Supreme Court has long encouraged law enforcement to obtain a warrant, rather than resorting to the alternative method of warrantless entries.[41] Although the consent exception to the warrant requirement provides an avenue for officers to avoid having to obtain a warrant, this exception is strictly scrutinised in its application as it involves the circumvention of a constitutional right.

As mentioned previously, the consent exception to the warrant requirement is valid only when consent is given voluntarily,[42] and consent is considered voluntary if it is the result of an "essentially free and unconstrained choice".[43] The factors to be considered by courts to determine the nature of the consent have been discussed in Part II above.

Courts have made it clear that consent cannot be considered voluntary when it is the product of coercion.[44] Physical interaction is not required for coercion to

[39] *US v Simmons* 172 F.3d 775 (Eleventh Circuit 1999) 778; *US v Valdez* 931 F.2d 1448 (Eleventh Circuit 1991) 1451–1452; *US v Garcia* 890 F.2d 355 (Eleventh Circuit 1989) 359–360 n 5; *Spivey* (n 1) 1219 ("Although voluntariness is usually a question of fact, the parties do not dispute the facts and both rely on the testimony of the government's witnesses.").

[40] *Kentucky v King* 563 US 452 (2011) 459; *Kyllo* (n 5) 31.

[41] *Ornelas v US* 517 US 690 (1996) 699.

[42] *Illinois v Rodriguez* 497 US 177 (1990) 181.

[43] *Schneckloth* (n 8) 225.

[44] *Bumper v North Carolina* 391 US 543 (1968) 550 ("Where there is coercion there cannot be consent."). See also *Schneckloth* (n 8) 227, 248 (stating that consent may not be "the result of duress or coercion, express or implied").

be present; mental coercion may also satisfy the threshold in certain situations.[45] As will be shown in this Part, there is considerable precedent supporting the notion that some instances of deception by law enforcement officers may amount to mental coercion. The Supreme Court has made it clear that when coercion is present there cannot be consent.[46] What remains unclear, however, is what degree of deception is necessary to constitute coercion.[47]

The exclusionary rule embraces the principle that the government must sometimes forfeit illegally obtained evidence, however incriminating it may be, to uphold the public's right to be free from unreasonable searches and seizures. In other words, although evidence may clearly incriminate a person, the rights of the guilty are upheld over the detriment of the evidence if it was illegally obtained. The rationale for the exclusionary rule is rooted in protecting the innocent from unreasonable intrusion when it is not certain that a person is guilty prior to the invasion. As Judge Cardozo (as he then was) said, under our exclusionary doctrine "the criminal is to go free because the constable has blundered".[48]

In the case of *Spivey*, there were two major misrepresentations that may have hindered Austin's ability to grant police the voluntary consent required to conduct a warrantless search. The first major misrepresentation used by the police to induce Austin's consent involved deceit of authority. It is clear the officers deceived Austin of their actual positions within the government organisations and, in doing so, misrepresented their authority while still acting under the colour of a government official. This case must be evaluated under the rules that govern officials acting under disclosure of their governmental authority. In other words, although Agent Lanfersiek disguised himself as a police officer, he was not considered to have been working 'undercover' given that the police are government employees in which the public places a heightened level of trust. The second critical misrepresentation conveyed by the officers in *Spivey* was deceit regarding the true purpose of their investigation. For subjects to give knowing and voluntary consent, they must understand the extent of what they are consenting to. This includes knowledge

[45] *Blackburn* (n 14) 206 ("the efficiency of the rack and thumbscrew can be matched, given the proper subject, by more sophisticated modes of 'persuasion'").

[46] *Bumper* (n 44). See also Schneckloth (n 8) 228 ("For, no matter how subtly the coercion was applied, the resulting 'consent' would be no more than a pretext for the unjustified police intrusion against which the Fourth Amendment is directed.").

[47] William E Underwood, '*A Little White Lie: The Dangers of Allowing Police Officers to Stretch the Truth As a Means to Gain a Suspect's Consent to Search*' (2011) 18 Washington & Lee J Civil Rts & Soc Just 167, 179 ("It is undisputed that valid consent must be freely and voluntarily given, but it is decidedly unclear what degree of falsehood is necessary to constitute outright coercion.").

[48] *People v Defore* 242 New York 13 (NY 1926) 21. See also Arnold H Loewy, 'The Fourth Amendment as a Device for Protecting the Innocent' (1983) 81 Michigan L Rev 1229 (Fourth Amendment can be invoked by the guilty "when necessary to protect the innocent").

of whether they are being aided by the investigation; constitute the targets of it; or both.

A. Deceit of authority

Prior to Spivey, the Eleventh Circuit had addressed the issue of police deception to induce consent. In *US v Tweel*, the Court held that consent searches are generally unreasonable when government agents acting under the colour of authority induce consent by "deceit, trickery, or misrepresentation".[49] The Court in *Tweel* went on to say that the Internal Revenue Service's (IRS) failure to disclose that they were investigating the defendant on the behalf of the Organised Crime and Racketeering Section of the Department of Justice, when asked if there was a special agent involved in the investigation, made the investigation "a sneaky deliberate deception" which rendered the defendant's consent involuntary.[50] *Tweel* demonstrates that deliberate omissions to disclose authority are deceptive and cannot be tolerated because they are an abuse of the public's trust in law enforcement that violates the protections of the Fourth Amendment. In *Spivey*, the police devised a sly and deliberate deception by misrepresenting Agent Lanfersiek's real authority. Disguising Agent Lanfersiek as a police officer was deliberately done to prevent Austin from knowing his true position of authority as a Secret Service agent. Just as in *Tweel*, Austin was not aware she was under investigation because the officers chose to deliberately hide their authority in an effort to deceive Austin into consenting to a search.

Securities and Exchange Commission (SEC) v ESM Government Securities, another Eleventh Circuit case that addressed the issue of police deception, specifically details the heightened duty of government officers acting in an official capacity to behave in a manner that will maintain public trust and cooperation:[51]

> We believe that a private person has the right to expect that the government, when acting in its own name, will behave honorably. When a government agent presents himself to a private individual, and seeks that individual's cooperation based on his status as a government agent, the individual should be able to rely on the agent's representations. We think it clearly improper for a government agent to gain access to [evidence] which would otherwise be unavailable to

[49] *US v Tweel* 550 Federal Reporter 2d 297 (Fifth Circuit 1977) 299. In *Bonner v City of Prichard* 661 Federal Reporter 2d 1206 (Eleventh Circuit 1981) (*en banc*) 1209, the Eleventh Circuit adopted as binding precedent all decisions of the former Fifth Circuit handed down before 1 October 1981.

[50] *Tweel* (n 49) 299.

[51] *Securities and Exchange Commission (SEC) v ESM Government Securities* 645 Federal Reporter 2d 310 (Fifth Circuit 1981) 316. In *Bonner* (n 49) 1209, the Eleventh Circuit adopted as binding precedent all decisions of the former Fifth Circuit handed down before 1 October 1981.

> him by invoking the private individual's trust in his government, only to betray that trust. When that government agency then invokes the power of a court to gather the fruits of its deception, we hold that there is an abuse of process.[52]

It is clear then, under the precedent set forth in *SEC v ESM Government Securities*, the officers in *Spivey* acted inappropriately to induce consent from Austin. The government did not act honourably by disguising a special agent of the Secret Service as a police officer. They created an elaborate ruse to make Austin believe they were there to investigate a burglary that she was the victim of, thereby seeking her cooperation and consent through misrepresentation. Austin placed her trust in officers acting under their official capacity only to learn that her trust had been betrayed and she was no longer the victim but instead the target. The consent given to officers by Austin was clearly the product of manipulative police deception and any fruits procured from the subsequent search should be suppressed.

Consent obtained by a fraudulent or mistaken claim of authority is not taken as lightly by the Supreme Court as the *Spivey* Court of Appeals majority may lead one to believe. In *Bumper v North Carolina*, the Court held that a consent obtained through deceit of lawful authority to conduct the search is considered so coercive that it is determined to be involuntary and will usually lead to suppression of the fruits of the search.[53] Furthermore, if a person does affirmatively respond to the fraudulent or mistaken claim of authority, a court that interprets this affirmative response as consent will generally hold that it was coerced.[54] Although Agent Lanfersiek represented himself as a local police officer to 'dust' for fingerprints needed to catch a perpetrator who was already in custody, he had no legal authority to investigate a burglary. Agent Lanfersiek was a member of the United States Secret Service whose federal jurisdiction does not extend to the investigation of a state claim such as burglary of a home. Unlike Agent Lanfersiek, Detective Iwaskewycz did technically have the legal authority to investigate a state claim. Detective Iwaskewycz, however, was not there to investigate for burglary, but instead was a member of a task force especially trained to investigate for fraud. Additionally, the case had already been closed by the neighbouring police department. For

[52] ibid.

[53] *Bumper* (n 44) 548–550.

[54] William E Ringel, Justin D Franklin and Steven C Bell, *Searches & Seizures, Arrests and Confessions* (2nd edn, Clark Boardman Company 1979) s 9:21.

the officers to act in their official capacity and pretend to continue the burglary investigation was a clear misrepresentation in the context of an already closed case.

B. Deceit of purpose

The second significant misrepresentation by officers that affected Austin's ability to give knowing and voluntary consent is the deceit of purpose. The Supreme Court has held that blatantly false statements by officers acting under official capacity are analogous to coercion.[55] For example, in *Bumper v North Carolina*, the Court held when a law enforcement officer claims that he has the authority of a warrant to search a home, when in fact he does not, he is essentially stating that the suspect has no right to resist.[56] Not all lies, however, are as blatant and clear cut as the one in *Bumper*. At what point does a misrepresentation to a suspect become so potent that it renders a suspect's consent involuntary? Professor LaFave, a criminal procedure scholar known for his work regarding the Fourth Amendment search and seizure, attempted to add some measure to this elusive issue. According to LaFave, when a misrepresentation is so extreme that it deprives an individual of his ability to accurately evaluate the situation, the subsequent consent is considered involuntary.[57] How could a misrepresentation as to the purpose of the officers' entry into a person's private home not be considered an extreme misrepresentation? Clearly, the purpose of an officer's request to enter a home plays a significant role in the calculation by the suspect as to whether to consent to the officer's entry.

An entry into a home that is the product of a ruse where the suspect is informed that the person seeking entry is a government agent, but is misinformed as to the real purpose for which the agent is seeking entrance, cannot be justified by consent obtained through such circumstances. There is ample case law that supports this notion. The Ninth Circuit addressed a set of facts similar to those in *Spivey* in which federal narcotics agents, along with local law enforcement officers, knocked on a suspect's door and asked permission to investigate a fictitious robbery in order to gain access into the residence.[58] The court in that case noted the differences between undercover officers and officers acting under the colour of office.[59] While deception is permissible in undercover work, the court expressed its disapproval of the latter misrepresenting their purpose while seeking entrance into a residence.[60] *Spivey* is comparable to *Bosse* in that federal agents accompanied

[55] *Bumper* (n 44) 550.

[56] ibid.

[57] Wayne R LaFave, Jerold H Israel & Nancy J King, *Criminal Procedure* (3rd edn, 2000) s 3.10(c) (stating that when consent is induced by extreme misrepresentations made by law enforcement officers acting in their official capacity, then the consent is rendered invalid).

[58] *US v Bosse*, 898 Federal Reporter 2d 113 (Ninth Circuit 1990) 115 (*per curiam*).

[59] ibid 116.

[60] ibid.

local law enforcement to Austin's home. Although they did not fabricate a robbery like the officers in *Bosse*, the officers did create an elaborate ruse to dress up and pretend to investigate a real burglary case, even though the case had been closed by the neighboring police department prior to their arrival at the residence. Agent Lanfersiek testified that he was not at the defendant's home to investigate a burglary. Instead, he was there to investigate them for fraud, a purpose which the officers clearly chose not to divulge to Austin and furthermore, officers created an elaborate ruse to avoid any disclosure of their real purpose for seeking entrance. Although Agent Lanfersiek was dressed as a local officer rather than wearing his normal uniform, both officers were acting under colour of their office, and not in an undercover capacity.

According to the Supreme Court in *Florida v Jardines*, the scope of a license to search—whether express or implicit—is limited to a particular area as well as *a specific purpose*.[61] At times, officers use deception to probe into situations where an ambiguity related to the crime still exists.[62] In *Jardines*, officers walked drug-sniffing dogs around a house to see if drugs were being grown in the house. When the dogs alerted the officers to the presence of drugs, the officers used that information to obtain a warrant to search the house. The Supreme Court suppressed the fruits of that search because the curtilage of a house is a constitutionally protected area in which the owner must give officers leave to be there. Although a front porch may concede an implied license for people to come knock, it does not grant officers permission to use a police dog to gather information through means beyond those expected of a person who enters the porch to knock on the door. According to the Court, so long as officers entered the constitutionally protected area for a purpose beyond the implied license, the physical intrusion constituted an unlicensed search of the premises. Similar to *Jardines*, the officers in *Spivey* went beyond the scope of Austin's license to search. If officers portrayed to Austin that the purpose of their entry was to further investigate the robbery, then they exceeded the scope of her license when they deliberately entered the home for the purpose of probing for evidence of credit card fraud.

When officers misrepresent their purpose for seeking entry into a residence, they capture that person's trust in law enforcement, only to subsequently betray it. The nature of the investigation that officers are there to conduct is a key ingredient of a subject's consent. When officers actively misrepresent the purpose of their investigation while acting in official capacity, they introduce an element of coercion

[61] *Florida v Jardines* 133 Supreme Court 1409 (2013) 1417.

[62] Elizabeth E Joh, 'Bait, Mask, and Ruse: Technology and Police Deception' (2015) 128 Harvard L Rev 246, 251.

that renders the consent involuntary.[63] As discussed above, there is ample case law aimed at discouraging such exploitation of a citizen's trust in the government in an effort to induce consent to search. The decision in *Spivey*, however, sanctions the use of such widely discouraged police deception to conceal the nature of an investigation. As will be observed, such a holding may give rise to policy concerns and consequential effects on the public's trust in law enforcement, and in the government, to protect their constitutional rights.

C. Deception material to consent

When officers misrepresent their authority and purpose for seeking entry into a home, they introduce an element of deception that may be critical to a subject's calculation of whether to grant consent. When police deception inhibits a suspect's ability to make a rational choice regarding consent, this deception may amount to coercion that overbears a defendant's free will.[64]

Although the majority downplayed the elaborate ruse concocted by the officers in *Spivey* as 'silly' at most, a further evaluation of the facts and surrounding circumstances of the case suggests that it played a material role in Austin's decision to consent. The majority in *Spivey* suggests that Austin's decision to consent would have been unaffected by the knowledge that Agent Lanfersiek was not a local police officer but a US Secret Service Agent specialising in fraud investigations. According to the majority, this intricate plan concocted by a group of about ten law enforcement officers was nothing more than a trifling detail in the case. However, the majority has significantly moderated the impact that this had on Austin's consent. What could have been more material to Austin than knowing the officers' true identities and the real nature and target of their investigation? If this fact of the case was so trivial, then why would officers have felt the need to devise such an elaborate ploy in the first place?

The answer is simple: this ruse amounted to a material factor in Austin's consent which amounted to coercion. Not only does the complexity of the ploy

[63] The court in *Washington v McCrorey* 851 Pacific Reporter 2d 1234 (WA 1993) 1240 held the following:

> It is improper for a government agent to gain entry by invoking the occupant's trust, then subsequently betraying that trust. Members of the public should be able to safely rely on the representations of government agents acting in their official capacity. We conclude that police acting in their official capacity may not actively misrepresent their purpose to gain entry or exceed the scope of consent given.

US v Turpin 707 Federal Reporter 2d 332 (Eighth Circuit 1983) 332–335 (misrepresentation about nature of investigation may be evidence of coercion).

[64] *US v Rutledge* 900 Federal Reporter 2d 1127 (Seventh Circuit 1990) 1129 (reviewing the voluntariness of a confession based on whether "the government has made it impossible for the defendant to make a rational choice as to whether to confess—has made it in other words impossible for him to weigh the pros and cons of confessing and go with the balance as it appears at the time").

itself indicate the importance it played in inducing Austin's consent, but the facts of the case also lend support to the idea that, absent such misrepresentations, Austin would not have given consent. A suspect's bewilderment upon learning they are the target of an investigation suggests lack of consent.[65] Austin's cooperation immediately shifted when Detective Iwaskewycz informed her of the real reason for their presence at the residence. When Austin became aware that she was no longer the 'victim' and instead was the 'suspect', she ceased to be a willing participant in the investigation, showing that the officers' deception was material to her decision to consent to their entry and cooperate in their investigation.

D. Circumventing the constitution

Allowing officers to arbitrarily circumvent the Fourth Amendment warrant requirement through strategy and deceit contradicts the purpose of the Fourth Amendment protections.[66] The protection of the warrant requirement calls for a neutral magistrate to make the inferences of whether the evidence provided by officers justifies intrusion into a suspect's home.[67] Without requiring a neutral magistrate to make such inferences, the protections provided by such requirement would be nullified, leaving the security of a person's home at the discretion of police officers.[68] In *Johnson v United States*, the Supreme Court made it clear that, as a default, when the right of privacy must reasonably submit to the power of search, a neutral judicial officer should make the determination of a power to search, not an officer who is acting as an interested party in ferreting out crime.[69]

There are, of course, instances where obtaining a warrant from a judicial constable may not always be reasonable or practical; but in such cases officers are required to show that these exceptional circumstances prevented their ability to obtain a warrant.[70] The case of *Spivey* is very similar to *Johnson* in that officers had no valid reason as to why they could not obtain a warrant. Although in *Johnson* the Court held that the suspect never consented to the search, the Court made it clear that the default rule is to obtain a search warrant when it is reasonable to do

[65] *US v Cabrera* 117 Federal Supplement 2d 1152 (KS 2000) 1159.

[66] Rebecca Strauss, 'We Can Do This the Easy Way or the Hard Way: The Use of Deceit to Induce Consent Searches' (2002) 100 Michigan L Rev 868, 875 ("The general intent of the Fourth Amendment was to limit the discretion and abuse of discretion by law enforcement to invade the privacy of citizens.").

[67] *Johnson v United States* 333 US 10 (1948) 14.

[68] ibid.

[69] ibid.

[70] ibid 14–15.

so and to resort to exceptions of the warrant requirement only when extenuating circumstances require such action.

In *Spivey*, there was clear evidence that the officers concocted this elaborate ruse specifically to circumvent the Fourth Amendment's warrant requirement. Furthermore, the officers did not demonstrate any valid reason for not obtaining a search warrant other than the inconvenience to the officers and the lack of evidence necessary to support the probable cause required for obtaining a warrant. As the Court concluded in *Johnson*, such reasons are not enough to justify bypassing the constitutional requirement.

The Supreme Court has warned that courts must be wary of police planning around constitutional protections.[71] The Eleventh Circuit has addressed the issue of police devising premeditated ruses to avoid constitutional requirements by stating that they refuse to "allow the state to secure by stratagem what the Fourth Amendment requires a warrant to produce".[72]

Warrant exceptions were created to provide relief when obtaining a warrant was not reasonable. Officers should not be allowed to exploit these exceptions by manipulating situations so that they may fit into one of the immunities. Out of all the exceptions that courts have carved out of the warrant requirement, the consent search is the most common type of warrantless search utilised by law enforcement.[73] This shows that officers commonly rely on this exception demonstrating the need for courts to establish guidelines to limit this strategic avoidance.

E. 'Curing' consent

The District Court in *Spivey* suggests that consent can later be 'cured' by the signing of a waiver to search the premises. However, there is abundant case law that rebuts this notion. For consent after an illegal seizure to be valid, the government must establish both the voluntariness of the consent to search and that the consent was not the product of illegal seizure.[74] Furthermore, a defendant's consent does not by itself cure the taint of an unlawful search. [75] An after-the-fact written consent may lead to an atmosphere of obligatory cooperation, as it is akin to a submission to authority. This is because the officers have already conducted the search and subjects are made aware that officers have likely already found any

[71] *Missouri v Seibert*, 542 US 600 (2004) 617 (holding that "[s]trategists dedicated to draining the substance out of constitutional protections cannot accomplish by planning around these protections because it 'effectively threatens to thwart [their] purpose'").

[72] *Graves v Beto* 424 Federal Reporter 524 (Fifth Circuit 1970) 525.

[73] Jones (n 2).

[74] US Constitution Amendment 4; *US v Hernandez-Penazola* 899 Federal Supplement 2d 1269 (Middle District of Florida 2012), appeal dismissed (Eleventh Circuit 2012).

[75] *US v Roberts* 888 Federal Supplement 2d 1316 (Northern District of Georgia 2012) 1324.

incriminating evidence that may be in the suspect's home.[76] The suspect's further cooperation may be an effort to show good faith. Cooperation that may lead to sympathetic prosecution rather than serve as evidence shows that the subject would have consented anyway prior to the search.

The officers' entry into the home was made for the purpose of conducting an undisclosed search aimed at exploiting the illegal activity of Austin and Spivey. The written waiver for a search warrant, which officers obtained from Spivey after a search of the home had already been done and incriminating evidence of their wrongdoing had already been procured, is not considered voluntary waiver. The officers deliberately exploited the evidence attained in the initial illegal search when they made the defendants aware of the evidence they already saw prior to obtaining the written waiver. In *Wong Sun*, the Supreme Court noted that evidence obtained through a lawless search in which officers then exploit to induce an after-the-fact consent of the search, does not dissipate the illegal taint of the initial search and the evidence remains 'fruit of the poisonous tree' and may not to be used in court.[77] As the warrantless search of the defendants' residence violates the Fourth Amendment, subsequent consents and subsequent statements are all 'fruit of the poisonous tree'. Officers used the contraband found in the initial search to exploit Spivey's consent to conduct a full-scale search.

F. Too vulnerable to be voluntary?

Another element that courts may consider when evaluating Austin's voluntariness based on the totality of the circumstances is vulnerability.[78] Austin and Spivey's residence had recently been burgled twice: a complete stranger had entered a place of sanctity and privacy for the defendants, had rummaged through their belongings and personal effects, and had taken things of value from them. A burglary not only deprives a victim of objects of value, but also robs them of their sense of security and privacy in their home. When Hunt burgled Austin and Spivey's residence twice, he left them feeling violated and vulnerable to outsiders.

When officers arrived at Austin's residence, her excitement in cooperating with the officers to help catch the burglar shows that she was in a position of vulnerability and was desperate to catch the person who had twice violated their sense of security in their home. She viewed the officers as authorities she could trust to aid her in restoring the sanctity of her home; but the officers betrayed this

[76] *US v Bushay* 859 Federal Supplement 2d 1335 (Northern District of Georgia 2012) 1346, 1353.

[77] *Wong Sun v US* 371 US 471 (1963) 488.

[78] *US v Parsons* 599 Federal Supplement 2d 592, 607 (noting that Parson's advanced age, poor physical and mental condition, and current living situation left him in a particularly vulnerable state).

trust and exploited her vulnerable state by using her desperation to induce consent to a search completely unrelated to catching the burglar.

V. Potential Policy Consequences

The courts have left a grey area between how much deception is permissible by government officials and how much deception constitutes coercion. Public policy considerations expose the ramifications that may result if courts sanction the level of police deception officers used in *Spivey*.

A. Sanctioning abuse of the consent exception

The Fourth Amendment is meant to be a safeguard to protect citizens from warrantless searches. Nevertheless, as the courts continue to carve out exceptions to the warrant requirement, they chisel away at the protective barrier the Amendment guarantees. The exceptions delineated by the courts are meant to be used in cases where extenuating circumstances make it impractical or impossible to obtain a warrant but not in cases where obtaining a warrant is merely burdensome for officers. The consent exception has become the most common type of warrantless search used by law enforcement and therefore should be evaluated with strict scrutiny as to the necessity of employing it.[79] The constitutional protection against unreasonable searches and seizures widens or narrows depending on the difficulty or ease with which the prosecution is required to establish consent.[80]

The decision in *Spivey* teaches police that they do not need to procure a warrant as long as they are able to formulate a plan to strategically circumvent the warrant requirement through the abuse of the consent exception. By sanctioning this practice, the courts are allowing officers to strategically circumvent the two key limits and protections provided by warrants: the existence of probable cause and the scope of the actual search. Courts are essentially giving officers the discretion of whether to obtain a warrant to conduct a search and exposing the public to the dangers of consent searches that may be almost limitless in scope.[81] When faced with the option, many officers will attempt to conduct the search without having

[79] Richard Van Duizend, L Paul Sutton and Charlotte A Carter, '*The Search Warrant Process: Preconceptions, Perceptions, and Practices*' (National Center for State Courts 1984) 21 (finding that "some ninety-eight percent of searches police conduct without a warrant they conduct pursuant to consent").

[80] Wayne R LaFave, *Search and Seizure: A Treatise on the Fourth Amendment* (2nd edn, West Pub Co 1986) s 8.1 ("Because of the frequent reliance upon consent searches, it is apparent that the constitutional protection against unreasonable searches and seizures widens or narrows, depending upon the difficulty or ease with which the prosecution can establish such consent.").

[81] Strauss (n 66).

to go through the burdensome process of obtaining a warrant. This contradicts the intent of the Fourth Amendment to limit the discretion and abuse by law enforcement officers to violate citizens' privacy.[82]

Instead, the courts should be encouraging the use of warrants that allow neutral magistrates to evaluate the facts to determine if probable cause exists and limit the scope of the search. Allowing the liberal use of consent searches essentially permits causeless, groundless, and boundary-free searches which contradicts the protections provided in the Fourth Amendment. If officers are given the discretion to choose whether to get a warrant, like the court in *Spivey* allows, then officers will be less compelled to apply for a warrant and will be encouraged to come up with creative ways to avoid having to obtain a warrant. Warrants face the possibility of being rejected due to lack of probable cause and requires officers to complete burdensome paperwork and patiently await the magistrate judge's approval.

B. Undermining public trust in a government institution

As a government institution, the police are held to a higher standard of integrity than other organisations. The public looks up to the government to maintain order and provide security in their life. If the police are allowed to violate the protections of the Constitution, they betray the *quid pro quo* relationship established between the citizens of a country and their government. If the police expect citizens to cooperate in their efforts to maintain law and order, there is a correlative duty upon officers to act with integrity and honesty towards citizens.[83] As Justice Brandies said in his dissent in *Olmstead v United States*,

> Our Government is the potent, the omnipresent teacher. For good or for ill, it teaches the whole people by its example… If the Government becomes a lawbreaker, it breeds contempt for law; it invites every man to become a law unto himself; it invites anarchy.[84]

People have a right to expect that the government, when acting in an official capacity, will behave in an honourable manner. Therefore, it is improper to allow government agents to induce consent through exploitation of a citizen's trust in the institution, only to betray that trust.[85] Allowing methodical police deception, such as the ruse employed in *Spivey*, will serve to undermine the public's faith in the government in many ways. For instance, when a citizen reports a crime, they

[82] ibid.

[83] Milton Hirsch, '*Wyche v State: A Case Analysis*' (2008) 33 Nova L Rev 137, 149.

[84] *Olmstead v United States* 277 US 438 (1928) 485.

[85] *US v Piper* 681 Federal Supplement 833 (Middle District of Georgia 1988) 837.

expect that officers will legitimately follow up on their report. Understandably, the proceeding investigation may require officers to investigate the scene of the incident. Instances such as the one in *Spivey*, however, will create paranoia for citizens who may normally cooperate with officers, in fear that the officers are not actually following up on the crime they reported. Allowing officers to deceive people about the purpose and nature of their investigation makes it impossible for citizens to know the real intent of officers seeking to enter their residence. It is not good practice for courts to require citizens to specifically inquire about *all* the purposes of a police officer's investigation, because this fosters public scepticism in the police as an institution. In the current climate of heated dispute regarding police shootings, authorised police deception may further increase scepticism from an already distrustful public that is predisposed to doubt police authority.

Authorising police deception would create a dual role for officers who are required to testify in the courtroom. The same officers that acted dishonestly in inducing consent to search would then be expected to convey a sense of trustworthiness and integrity when testifying in court.[86]

C. Sanctioning deceptive police protocols

As mentioned previously, police protocol often encourages officers to use deceptive techniques and psychological manipulation when investigating a crime. As far as police are concerned, lies relating to police authority or purpose are often considered "techniques of the trade" that should be fostered rather than condemned.[87]

The courtroom is not the only place where officers face competing expectations. Officers are expected to function as both peacekeepers and investigators. The public demands a level of transparency with officers to ensure their abilities to keep them safe, but often officers are forced to conceal the truth and even lie about facts and circumstances to complete their duty as peacekeepers.[88] Furthermore, if officers are constantly fulfilling both roles, those who are seeking the assistance of the peacekeeper then become vulnerable subjects to the investigator.

Allowing officers to deceive citizens in an effort to fulfil this dual role does not come without consequences. If reporting a crime means that a person may become the subject of an investigation themselves, people may stop reporting crimes. Even the innocent may become paranoid of police deception and exploitation of their need for assistance that they are deterred from seeking police help for

[86] Jones (n 2) 530.

[87] Jerome H Skolnick, *Justice Without Trial: Law Enforcement in Democratic Society* (John Wiley & Sons Inc 1967) 196–197.

[88] Jones (n 2) 529.

fear of an investigation turning up some unintentional or trivial illegal activity. But the current legal framework, as highlighted by the majorities' opinion—that those caught in these broad webs of deception have assumed the risk—give officers nearly limitless discretion.[89] Since citizens will have no way of knowing whether their cooperation or assistance to law enforcement is being called on for public or personal good or for the purpose of incriminating them, they will be discouraged from aiding in the apprehension of criminals.[90] This decision essentially requires citizens to confront officers about the purpose of their investigation which may, in turn, signal to officers that the suspect is being uncooperative and may potentially trigger friction on both sides of the encounter.

Encouraging officers to employ deceptive techniques is an outcome-focused practice that is adverse to the traditional criminal justice paradigm that places the value and materiality of evidentiary fruit secondary to the legality of its procurement.[91] Police deception seems to run contrary to social instinct that usually commands honesty and transparency from those seeking cooperation from others.[92] In essence, it seems that the police are holding their societal function above the manners of social norms and, based on the value of evidence that such practices may unearth, hold their methods as an exception to the guiding principles of criminal justice.

VI. Drawing Some Lines on Police Deception

Courts have taken a *laissez-faire* approach to the issue of police deception as it pertains to constitutional rights for far too long. As new technologies develop and continue to shape the Fourth Amendment, lines need to be drawn on how far officers can go before deceptive techniques overbear a person's free will to consent. Courts must tread carefully when developing new approaches to interpreting the Fourth Amendment because they often require the courts to weigh the individual rights of citizens against crime control and societal welfare; too much lenience with officers may lead the people to think the courts are favouring a more tyrannical form of governmental control, whereas too much restriction on police may lead

89 Joh (n 62) 251.

90 *Krause v Commonwealth* 206 South Western Reporter 3d 922 (KY 2006) 926.

91 Jones (n 2) 529.

92 Irina Khasin, '*Honesty is the Best Policy: A Case for the Limitation of Deceptive Police Interrogation Practices in the United States*' (2009) 42 Vanderbilt Journal Transnational Law 1029, 1037 (noting that "the practice of police deception runs contrary not only to widely held beliefs about right and wrong, but also to the ideals of the American criminal justice system").

people to believe the courts are too soft on crime and may raise safety concerns for society.

Thus far, the Supreme Court has not addressed deception in the context of inducing consent to search.[93] According to LaFave, there is no communal understanding as to what constitutes permissible deception in criminal policing.[94] With few to no restrictions on police deception in obtaining consent to search, there remains a hollow void in the understanding of how far police may go to persuade a subject's consent. There are three rules proposed in this article to help fill in the gap on police deception and offer some guidance on creating a more structured analysis for courts to evaluate their decisions on police trickery.

The first rule proposed is to prohibit officers from impersonating other government officials in an effort to induce a subject's consent. When an officer impersonates another type of government agent, they are still acting under colour of governmental authority. In *Bumper*, the Supreme Court held that a consent to search obtained by a fraudulent claim of authority is considered coercive and involuntary and may lead to suppression of any fruits procured from the search.[95] Although *Bumper* involved a situation in which officers made a fraudulent claim that they had a warrant to search the residence when they actually did not have one, the rule derived from this case lends itself to the rule proposed in this section. Similar to the facts in *Bumper*, the officers in *Spivey* made a fraudulent claim of authority when they had Agent Lanfersiek disguise himself as a local police officer. Police officers have the authority to investigate burglaries, such as the one that they lead Austin to believe they were there to investigate. By acting under the guise of a local officer, Agent Lanfersiek portrayed to Austin that he had the authority to investigate a burglary when in fact he did not. This rule simply extends this prohibition on fraudulent claims of authority to induce consent to include officers going undercover as other government agents to disguise their lawful authority. As discussed throughout this article, when acting under the capacity of a government official, officers are expected to behave honourably. When these agents use the trust that citizens have in the government to provoke their consent, any deception used by them reflects poorly on the nobility of the institution and undermines the trust that people have in the government.

The second rule proposed is that there should be a 'meeting of the minds' on the purpose of the search that officers are seeking consent for. The plain view doctrine was not meant for officers to abuse by manipulating a subject in a way that will induce consent to enter with the hopes that officers will then find incriminating

[93] Strauss (n 66) 884.
[94] LaFave (n 80) s 8.2(n).
[95] *Bumper* (n 44) 548–550.

evidence once inside. By allowing officers to pretend they are at a residence for a specific purpose, whereas in fact the actual purpose remains undisclosed, crosses the line from a simple strategic omission to a deceptive manipulation meant to trick the suspect into giving their permission to search.

Officers acting in a disclosed capacity should be required to divulge their prior intentions while conducting investigation before obtaining consent to search from a suspect. If there is no meeting of the minds on the subject matter of a search that officers are seeking to conduct, the scope of the consent cannot be considered to authorise any search executed for a purpose beyond what the suspect was aware of at the time of consent. In *Florida v Jimeno*, the Court made it clear that the burden lies on the defendant to set express parameters to limit the scope of their consent.[96] It is not possible, however, to set boundaries on the purpose of consent when citizens are deceived as to the purpose of the investigation and therefore, citizens require greater protection in this area to safeguard their Fourth Amendment right.

Although this rule may seem simple on its face, applying it in a fashion that accords the Court's approach to evaluating Fourth Amendment violations may not be so easy. Prior case law maintains that the subjective intentions of police officers play no role in evaluating Fourth Amendment violations.[97] Also, dissecting the true nature of all investigations when there are multiple being conducted on one occasion proves difficult when officers are not willing to be candid about what their intentions were prior to searching the residence.

Courts should evaluate this second rule through an objective lens. To determine if there was in fact a meeting of the minds as to the purpose of the investigation that the subject consented to, courts should consider the totality of the circumstances of each case. When evaluating the totality of the circumstances there are a number of factors which courts may analyse to reach a determination.

One factor that the courts may find useful to consider is the nature of the officer's primary assignment within the government organisation at the time of the search. A specific assignment of an officer within an organisation will shed light

[96] *Florida v Jimeno* 500 US 248 (1991) 251.

[97] *Whren v United States* 517 US 806 (1996) 813 (noting that pretextual stops based on an officer's subjective motivations, if otherwise supported by probable cause, are permissible because courts only look objectively at whether there was probably cause or reasonable suspicion for the stop); *Horton v California* 496 US 128 (1990) 138 (stating that "evenhanded law enforcement practices are best achieved by the application of objective standards of conduct, rather than the subjective state of mind of an officer"); *Devenpeck v Alford* 543 US 146 (2004) 154–155 (holding that an arresting officer's subjective reason for making an arrest is irrelevant when the criminal offense as to which the known facts provide probable cause).

on their role in an investigation and the nature of the crime for which they seek to gain evidence. An assignment to a specific task force or agency may also reveal the authority of an officer to investigate specific crimes. For instance, Agent Lanfersiek was a member of the Secret Service assigned to a fraud task force. Therefore, his assignment reveals that his main intentions for being at the residence was to investigate Austin and Spivey for fraud and that as a federal agent, he had no authority to investigate local burglaries. Detective Iwaskewycz was also assigned to the fraud task force within the Lauderhill Police Department. Although he technically has the authority to investigate local burglaries, his division assignment channels his role within the department to fraud-related crimes.

Another factor that courts may find useful to consider is the nature of the evidence collected by the officers in connection to the suspect. Although the plain view doctrine allows officers to collect evidence of any type of crime they visually detect in a legal search, when you combine the evidence procured with the assignment within the organisation, this may lend additional support towards the true nature of the investigation. In *Spivey* the prints collected were clearly a sham because Agent Lanfersiek was unfamiliar with the practices of how to gather fingerprints. Aside from the tapes which could have easily been requested and retrieved without intrusion, the only real evidence procured from this search was evidence of credit card fraud to be used against the residents. When paired with their assignments, this shows that the main motivation for the search was to uncover evidence for credit card fraud by inducing consent in order to avoid having to obtain a warrant.

Another factor to be considered may be any deliberate investigative measures taken to investigate a crime that was not disclosed to the defendant prior to consent. Although there are no examples of this in the case of *Spivey*, this factor may be useful in a situation where officers enter a dwelling under certain claimed intentions but employ search methods inconsistent with the reason for entrance that they conveyed to the subject who granted consent.

Another element that the courts may find useful to consider is whether the subject is aware that anything officers see in the house may also be used against them. This is in no way a requirement for officers to advise all subjects of a search that anything found within the residence during a search may be used against them. This educational factor is to simply give credit to officers who intentionally help reduce the possibility for deception when they inform subjects of the potential consequences of their consent. Although the courts should not give this factor heavy weight in their analysis, if an officer makes the subject aware of the significance of

their consent, it should be considered as a factor weighing against abusive police deception.

The courts should weigh these factors, as well as any other factors presented in each case, to determine if the totality of the circumstances support a finding of a 'meeting of the minds' regarding a subject's consent. It is again important to point out that this rule will apply only to officers requesting consent to search while acting in their official capacity.

The third rule, which really functions as an alternative to the second rule mentioned above, is to require officers to identify the task force they are acting under when requesting consent to conduct a search. Although courts tend to shy away from creating bright-line rules regarding the Fourth Amendment, crafting such a requirement would be useful in preventing major deceptive ruses such as the one presented in *Spivey*. Requiring officers to reveal their task force within the organisation generates more transparency between citizens and officers. If officers are not attempting to trick or deceive a subject as to the intentions of their search, they should have no problem disclosing their particular task force assignments to the subjects for which they are seeking consent. For example, in *Spivey*, if the officers had disclosed their assignments prior to requesting consent from Austin, this would dramatically reduce the potential for overwhelming police deception. This would mean that Austin was aware that the officers were part of the fraud task force and that there is the potential that they may be investigating her for fraud as well. It allows for the subject to delineate the intentions of the officers for herself and make a more informed decision of consent based on her evaluation.

A. Benefits to citizens

When crafting a new guiding principle in the law it is important for courts to weigh the interests of the government against the interest of the citizens before implementing the new principle. The citizens are in a position to receive the most benefit from any of the rules listed above. Implementing any of these rules will serve as a protective device on a citizen's Fourth Amendment rights. All of the rules are aimed at preventing officers from using twofold strategies in which they find some tenuous excuse to investigate a house to avoid disclosing their main purpose for being there. The first and third rules suggested are directed at preventing officers from taking their deception too far while acting under the colour of the law. The second rule is meant to prevent citizens from essentially consenting to a lie under the pretence that officers are there to aid them, when in fact they are the target of an investigation. Creating rules such as these will allow citizens to gain more trust in law enforcement, especially in a time of turmoil with law enforcement.

Without any guidance on police deception, police are free to deceive the public on what they are consenting to which may foster distrust in the government. This is especially important in situations such as the one in *Spivey* when officers attempt to strategise around the requirements provided in the Fourth Amendment.

B. Benefits to law enforcement

Implementing rules such as the ones suggested above will create clearer guidelines about the amount of deception police may use when seeking consent. By creating clearer guidelines, officers will not be as likely to risk doing something that may cause evidence that is vital to their case to be suppressed. Although bright-line rules are not generally used in regards to the Fourth Amendment, police need clearer guidance on just how far they may take their deception when acting under the colour of authority. Officers have many technicalities they must abide by to prevent getting their cases thrown out. These suggestions will provide officers with clear rules to abide by without having to decipher the grey area of law that surrounds police deception.

C. Benefits to courts

Applying guidelines such as the ones suggested above will further benefit the courts by creating a structure in which they may analyse whether police deception has crossed the constitutional boundaries provided by the Fourth Amendment. The second suggested rule allows the courts to weigh the costs of police deception with the benefits based on the amount of deception used to obtain consent, without having to draw a bright-line rule. Implementing guidelines for courts to evaluate cases of deception will provide consistency in application while preventing officers from having free-range discretion on how much deception they may use to induce consent.

Overall, these rules promote the honesty and integrity necessary to maintain a certain level of trust between law enforcement and the public when operating under the colour of government authority. Fostering and preserving the public's faith and trust in government agencies is essential to maintaining the public's cooperation with the institution.

VII. Conclusion

Consent to search that is given to officers who have deliberately misrepresented their authority and intent for their presence in order to circumvent the need for a warrant is not voluntary consent and violates the Fourth Amendment. According to Justice Scalia, the warrant requirement in the Fourth Amendment has become

"so riddled with exceptions that it [is] basically unrecognizable".[98] To alleviate some of the confusion caused by all of these exceptions, the Court needs to implement guidelines that draw some clear parameters by which officers should abide. The current judicial framework on police deception encourages deceitful consent searches, despite the consequential impact this may have on privacy and the contradiction it has with the purposes of the Fourth Amendment. The courts should adopt rules to help provide parameters for police use of deception to obtain consent to search. Implementing such rules will help to maintain the trust needed between citizens and law enforcement and help provide procedural guidelines to officers who wish to seize evidence vital to their case without fear that it will be suppressed.

[98] *California v Acevedo* 500 US 565 (1991) 582 (concurrence).