CAMBRIDGE LAW REVIEW

VOLUME I
2016

Editors-In-Chief
Ruth McGuinness & Peter Bruland

Proudly Supported By
Cambridge University Law Society
Slaughter and May

CAMBRIDGE LAW REVIEW

VOLUME I
2016

Editors-In-Chief
Ruth McGuinness & Peter Bruland

Proudly Supported By
Cambridge University Law Society
Slaughter and May

Cambridge Law Review

Typeset in Baskerville
Printed & Bound in the United Kingdom
for Cambridge Law Society
by Triple Take Ltd, www.tripletake.co.uk

Table of Contents

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

Ruth McGuinness & Peter Bruland

Despite their predominance and influence elsewhere in the common law world, student-published law reviews are relative newcomers to British academe. This scarcity of student-run publications does the academic community a disservice. Not only do such journals provide unparalleled opportunities for student editors, but they also occupy a special place within legal discourse, offering a unique degree of independence, editorial involvement and dexterity in responding to recent events. Recognizing the important contributions to be made by student-run legal publications, it is with great pleasure that we present the inaugural Volume of the Cambridge Law Review.

Founded to provide a forum for legal scholars, practitioners, and current students to contribute to academic debate within law, the Cambridge Law Review is edited by a rigorously selected group of graduate and undergraduate students at the University of Cambridge Faculty of Law. We owe heartfelt thanks to each of these talented editors, and especially to our vice editors-in-chief: Emily Gordon, Vicki Halsall, Oriyan Prizant and Andreas Wildner. In addition to the tireless work of our staff, this Volume was also made possible by the support of the Cambridge University Law Society and by the generous sponsorship of Slaughter and May. We are deeply grateful to them and to Craig Slade at Triple Take for helping us to make the Cambridge Law Review a reality.

In its first year, the Review attracted a multitude of excellent submissions from jurisdictions around the world. This Volume offers a diverse collection of articles, covering an array of topics of interest to both British and international audiences. These range from particularly timely analyses of European financial regulation in the aftermath of Brexit and police cautioning in England and Wales, to a case comment on a 2015 religious liberty decision by the Supreme Court of Canada, to an examination of the twenty-first century challenges facing United Nations peacekeeping missions. The scope and depth of these articles is a credit to the promising authors whose scholarship we are proud to publish.

It has been our pleasure and privilege to edit this inaugural Volume, and we wish the incoming Editorial Board every success with Volume 2. As student-run legal publications grow more common across the United Kingdom, we are confident that the Cambridge Law Review will take its place at the forefront of legal scholarship.

Ruth McGuinness and Peter Bruland
Founders and Editors-in-Chief

Note From Our Sponsors Slaughter and May

Slaughter and May is delighted to be associated with the inaugural volume of Cambridge Law Review.

Law reviews are an invaluable resource for legal scholars and practitioners the world over; they synthesise law reports, provide insight into judicial decisions, commentate on statute and contribute to the academic development of the law.

In commercial law, where we practise, the rules emanating from cases such as Carlill v. Carbolic Smoke Ball Co, Donoghue v. Stevenson and Caparo Industries v. Dickman (to name just three) are applied on a daily basis: they guide how contracts are drafted and they form the basis of advice to clients, whether or not they are cited by name!

Some of these rules are old and well loved, and the way they adapt and change to the modern, increasingly globalised economy is of vital importance to the practitioner of commercial law. Law reviews, such as Cambridge Law Review, provide an invaluable insight into how the law is developing and may expect to develop.

But it is not just about commercial law: the articles in this journal range from developments in the internet transmission of television and the relationship between privacy and data protection, on the one hand, to the legal hazards of mosh pits and how to ensure equality for women with mental health issues, on the other.

All of them are excellent, and it is to the credit of Cambridge University Law Society that it has attracted such diverse and quality submissions.

We hope you enjoy Volume 1.

Guy O'Keefe
Partner

Slaughter and May
25th September 2016

Note From The President Of Cambridge University Law Society

The publication of Volume 1 of the Cambridge Law Review marks a new high point in the history of the 115-year-old Cambridge University Law Society. It comes at an incredibly exciting time as the Society continues to expand. We are offering our members, and also the wider legal community, ever more resources and opportunities to support engagement with the law. It is thus a privilege to be the President overseeing the inception of Cambridge Law Review and to be asked to write an introductory note.

The breadth of content in the journal is to be commended. The contributors span the globe, and the legal issues discussed germane for a 21st Century student publication. It is all at once both informative and highly thought provoking. From Brexit related discussions to Mosh Pits and criminal liability, Volume 1 has set a set a high standard for this journal

Most importantly, the Cambridge Law Review simply would not have happened had it not been for the talent and energy brought by the Editors, Ruth and Peter. On behalf of the Society I would like to extend a big thank you to them. I have seen first-hand the time and effort put into building up the infrastructure to get Cambridge Law Review off the ground. Their passion and dedication has made this project a success. It has been a joy to work with you. I would also like to thank, on behalf of the Society, the Vice-Editors and the Editorial Board. The breadth, depth and quality of the articles included in this journal are testament to your tenacity and dexterity on this project. Lastly, I would to thank Slaughter and May for their tremendous help in financing the Cambridge Law Review.

The benchmark for this journal has been set, and I am very excited for its future.

Jack Lewis
President of Cambridge University Law Society

The White & Carter Legitimate Interest Qualification On The Elective Theory Of Contractual Repudiations: A Reformulation Proposal

Oriyan Prizant[1]

1. Introduction

UNDER ENGLISH CONTRACT Law, a party to a contract faced with a repudiation by the other can choose whether to accept the repudiation and treat the contract as terminated, or reject it and treat the contract as subsisting.[2] This is sometimes referred to as the 'Elective Theory'.[3] In *White & Carter v McGregor*,[4] Lord Reid, speaking in the House of Lords, introduced two qualifications on this ability to choose, the 'cooperation qualification' and the requirement that the rejecting party have a 'legitimate interest' in actual performance.[5] However, in enunciating the latter qualification, Lord Reid failed to provide a sufficiently clear definition of 'legitimate'. Lord Reid held that a purely financial interest can be legitimate without sufficiently clarifying what was meant by 'financial'. This ambiguity has resulted in uncertain judgments, and, most recently, an unnecessary invocation of the 'good faith' doctrine in *MSC Mediterranean v Cottonex Anstalt*[6] to resolve the ambiguity and find the financial interest claimed to be illegitimate.

This article seeks to demonstrate that the 'legitimate interest' qualification requires clarification. This article will first examine the state of the 'legitimate

[1] Bachelor of Law, Wolfson College, Cambridge. I would like to thank Dr. Janet O'Sullivan for her guidance and input; and Mr. Hananel Levi for his comments. In addition, I would like to thank my fellow editors of the *Cambridge Law Review*. All errors are entirely my own.

[2] *Howard v Pickford Tool Co Ltd* [1951] 1 KB 417, 421.

[3] *London Transport Executive v Clarke* [1981] ICR 355, 367, per Templeman LJ.

[4] *White & Carter (Councils) Ltd v McGregor* [1962] AC 413.

[5] ibid 431.

[6] [2015] 2 All ER (Comm) 614.

interest' qualification prior to the *MSC Mediterranean* decision. I will then analyse what changes, if any, Legatt J's judgment introduces into the current definition of 'legitimate interest'. Finally, I will then suggest a more appropriate construction of the 'legitimate interest' qualification and consider it against the 'reasonableness' alternative offered by some judgments.[7]

2. Development of the Law on the 'Legitimate Interest' Qualification

In *White & Carter*,[8] a garage and an advertising company had a three-year contract for advertising the former on the latter's dustbins. Near the end of the three-year contract, a manager at the garage agreed to renew it, though he had no authority to do so. The garage quickly informed the advertiser that it was an unauthorised renewal and that they did not intend to honour the new contract. The advertiser decided to proceed with performance of the contract, hence rejecting the repudiation. The contract included an accelerated payment clause for the entire three years, which was triggered when the garage failed to pay the monthly installments due under the renewed contract. The House of Lords affirmed that the innocent party has a right to elect to reject or accept a repudiation, even though it may lead to harsh results. The advertisers were awarded the entire three years' worth of installments owed under the payment acceleration clause. In handing down the leading judgement, Lord Reid held that this right of the innocent party may be qualified in two ways. The first of which is:

> …if it can be shown that a person has no legitimate interest, financial or otherwise, in performing the contract rather than claiming damages, he ought not to be allowed to saddle the other party with an additional burden with no benefit to himself. If a party has no interest to enforce a stipulation, he cannot in general enforce it: so it might be said that, if a party has no interest to insist on a particular remedy, he ought not to be allowed to insist on it.[9]

Lord Reid's formulation seems to indicate that a lack of legitimate interest will be found when two circumstances are met; firstly, where damages are an appropriate financial remedy for the rejecting party's losses, and secondly, where the rejecting party gains no additional benefit from the performance of the contract to the financial one securable by damages.

The characterisation of Lord Reid's judgment as placing qualifications on the innocent party's ability to choose was strongly rejected by Professor

[7] *Stocznia Gdanska SA v Latvian Shipping Co* [1998] 1 WLR. 574.
[8] *White & Carter* (n 4) 413.
[9] ibid 431.

Andrew Burrows[10] in his case note on *Société Générale London Branch v Geys*,[11] in which he criticises Lord Sumption's dissent due to its deployment of Lord Reid's qualifications in support of the 'automatic' termination theory.[12] In *Geys*, Lord Sumption stressed that the orthodox interpretation of the *White & Carter* decision places two qualifications on the innocent party's right to reject a repudiation. Lord Sumption relied on the 'co-operation' qualification to hold that an employee cannot be considered to have meaningfully rejected a repudiation when the contract is so reliant on the mutual relationship inherent in contracts of employment:

> If Lord Reid's qualifications to this proposition are ignored, this unattractive consequence will be gratuitously extended, at least in the context of contracts of employment, to cases where there can be no contractual performance, because the relationship is dead and all that survives is the husk or shell of a contract devoid of practical content.[13]

Lord Sumption goes on to hold that in such circumstances, the contract must be considered to have terminated, irrespective of the employee's rejection of the repudiation, by operation of Lord Reid's second qualification.

Burrows rejects Lord Sumption's dissent as he views it as a 'novel' interpretation of Lord Reid's judgment.[14] Burrows also points out that, for the specific employment law context, Lord Sumption placed too much weight on the disputed notion that unpaid wages cannot be claimed as debt.[15] Though a strong critique, it nevertheless misses the mark. Characterising the *White & Carter* decision as placing qualifications on the innocent party's ability to accept or reject a repudiation is the proper analysis for two principal reasons. Firstly, when discussing the 'legitimate interest' question, Lord Reid clearly phrases himself so to place a limitation on the otherwise uninhibited right to accept or reject a repudiation or a repudiatory breach. Lord Reid holds that lack of legitimate interest 'can be shown' and if this burden of proof is managed, the other party 'ought not to be allowed to insist on [the illegitimate interest].'[16] This was later applied in *The Alaskan Trader* to preven the rejecting party from doing so due to the burden of proving illegitimacy of the interest being met.[17] This naturally qualifies the previously unrestricted ability to elect. Secondly, the courts' subsequent usage of those limitations has clearly cemented their status as 'qualifications'. The manner in which the claims and

[10] Andrew Burrows, '*What is the effect of a repudiatory breach of a contract of employment*' (2013) 42(3) ILJ.
[11] [2013] 1 AC 523.
[12] ibid at [111].
[13] *Geys,* (n 11), 578.
[14] Burrows, (n 10), 287.
[15] ibid.
[16] *White & Carter* (n 4) 431.
[17] [1984] 1 All E.R. 129.

defences are presented to the court requires them to consider these as limitations to be proven or disproven.[18] In addition, the courts refer to Lord Reid's dictum as placing limitations on the right to elect. Most notably, in *The Odenfield* case, Kerr J referred to Lord Reid's dictum as placing 'fetters' on the unfettered right of the innocent party to elect whether or not to accept the repudiation.[19]

Burrows' main issue with Lord Sumption's analysis is that it treats the second qualification as advancing an automatic theory of termination in the specific case. To resolve this, it must be understood that the two qualifications are different in nature. The 'legitimate interest' one is a *legal* qualification assessed and inquired into by the courts. The 'cooperation' qualification, on the other hand, is an economic reality limitation which is *factual* in nature; it is either present or it is not. The latter simply dismisses sterile rejections which have no effect as the rejecting party cannot perform in any case. In the employment context, making oneself available and willing to work is sufficient to amount to performance. In this way, the worry of the automatic theory re-emerging via acceptance of Lord Sumption's logical analysis is clearly dispelled. Accepting the qualification analysis does not require acceptance of the automatic theory.

Lord Reid's 'legitimate interest' qualification was held to have been correctly deployed by the arbitrator in *The Alaskan Trader.*[20] The case concerned a two-year charter-party contract, twelve months into which the vessel required major repair work. A month into the repair work, the charterers communicated repudiation of the contract to the ship-owners, which the latter rejected. The repair work took eight months, at the end of which the ship-owner kept the vessel docked and fully staffed for the repudiating charterers for the remaining five months under the contract. The arbitrator found that the ship-owners' financial interest could have been properly met by way of damages and that it had no *additional* interest (on top of the financial one) in the performance of the contract. The arbitrator then held that the ship-owner, owing to its lack of legitimate interest to keep the contract alive, had to accept the repudiation when made, and could not claim in debt for the 5 months in which he kept the ship at the ready. Bingham J affirmed this finding as he found no flaw in the logic of the experienced commercial arbitrator.

This is a very sensible formulation. The damages measure, though not extensive under English law due to the requirement on the harmed party to mitigate its losses,[21] should in theory put a party in the position it would have been at had the contract been performed.[22] There is therefore no incentive to perform the contract for the sake of financial gain. If advertisers or ship-owners secure the full measure of damages they are compensated to the level of an equally

[18] Reichman and another v Beveridge and another [2007] Bus. L.R. 412, [15].
[19] *Gator Shipping Corporation v Trans-Asiatic Oil Ltd SA (The Odenfeld)* [1978] 2 Lloyd's Rep 357, 374.
[20] *The Alaskan Trader (No.2)* [1984] 1 All E.R. 129.
[21] *British Westinghouse Electric Co Ltd v Underground Electric Railways Co of London* [1912] AC 673.
[22] *Robinson v Harman (1848)* 1 Exch. 850, 855 per Parke B.

financially beneficial contract.[23] If losses are mitigated only halfway, or a reasonable failed attempt to mitigate is made, the damages measure rounds up their losses. Furthermore, losing parties should receive financial compensation to account for their efforts in attempting to locate a new contract to replace it (the mitigation). Thus, if the innocent party cannot enter alternative ventures of equal duration or benefit, the damages it will receive for the repudiated contract will cover that gap of lucrativeness. The original contract is therefore 'performed' financially in any case.

However, this interpretation was not exactly shared by the other cases on this topic. In *The Puerto Buitrago*,[24] a chartered ship required extensive and highly expensive repairs. The charterers repudiated the contract. The ship-owners rejected the repudiation and claimed that the charterers owed them the charter fee for the the time it would have taken to repair the ship. The standard imposed by the court of appeal for an interest being illegitimate was that of reasonableness. Lord Denning held that:

> ...the plaintiff ought, in all reason, to accept the repudiation and sue for damages—provided that damages would provide an adequate remedy for any loss suffered by him. The reason is because, by suing for the money, the plaintiff is seeking to enforce specific performance of the contract—and he should not be allowed to do so when damages would be an adequate remedy.[25]

Lord Denning's ruling indicates that if the party should 'in all reason' accept the repudiation it has no legitimate interest. The reasonableness language was subsequently tightened in *The Odenfield*.[26] Mr Justice Kerr held that a 'wholly unreasonable' standard applied to evaluate the legitimacy of the rejecting party's interest:

> any fetter on the innocent party's right of election whether or not to accept a repudiation will only be applied... where damage would be an adequate remedy and where an election to keep the contract alive would be wholly unreasonable.

The language used by Kerr J is akin to '*Wednesbury unreasonableness*'.[27] In other words, the party will not have a 'legitimate interest' if no reasonable contractual party in its position would elect to reject the repudiation. Kerr J used the language of 'wholly unreasonable, quite unrealistic, unreasonable and untenable'.[28] In

[23] ibid.
[24] *The Puerto Buitrago* [1976] 1 Lloyd's Rep. 250.
[25] ibid 259.
[26] *The Odenfeld* (n 19).
[27] *Associated Provincial Picture Houses Ltd. v Wednesbury Corporation* [1948] 1 KB 223.
[28] *The Odenfeld* (n 19).

Stocznia Gdanska SA v Latvian Shipping Co & Ors,[29] Clarke J found that the varying standards for the 'legitimate interest' qualification all pointed towards a general 'unreasonableness' test:

> Thus, on the footing that the principle exists, it is that an innocent party is entitled to continue to perform a commercial contract which has been repudiated by the other party unless he has 'no legitimate interest, financial or otherwise, in performing the contract' (per Lord Reid) or he should 'in all reason' accept the repudiation (per Lord Denning), B or where it would be 'wholly unreasonable' to keep the contract alive (per Kerr J) ... I do not think that there is any real difference between these differing ways of putting the principle. The question is therefore whether the buyer has an arguable case that the builder's decision... was wholly unreasonable.[30]

In *The Aquafaith,*[31] the most recent authority on the interpretation of the 'legitimate interest' qualification, Cooke J stated that the party rejecting a repudiation has a legitimate interest unless it is 'wholly unreasonable' or 'perverse', for it to complete performance.[32] This cements that the current interpretation of a 'legitimate interest' is virtually any interest that is not 'perverse', 'wholly unreasonable', or 'beyond all reason'; including a purely financial one.[33] In *The Aquafaith* Cooke J preferred the previously established standard of 'wholly unreasonable' conduct[34] for assessing legitimacy and found that, although mitigation and damages would secure an economically-equivalent result, a purely financial legitimate interest arose.[35]

3. *MSC Meditteranean v. Cottonex Anstalt:* the Need to Reformulate the Legitimate Interest Qualification

MSC Mediterranean v. Cottonex Anstalt concerned a carriage contract between company A, which provided containers for company B to transport its goods to be sold to company C in Bangladesh.[36] The sale of goods contract between B and C provided that title to the goods did not pass until full payment was made. C paid by way of letter of credit issued through its bank to B. However, once the

[29] [1995] CLC 956.
[30] ibid 968.
[31] [2012] EWHC 1077 (Comm).
[32] *Stocznia Gdanska* (n 25) 968.
[33] ibid 915.
[34] *The Puerto Buitrago* (n 24); *(The Odenfeld)* (n 19).
[35] *The Aquafaith* (n 31) 909–910.
[36] [2015] EWHC 283 (Comm).

containers arrived at Bangladesh, there was a sharp fall in the price of cotton. C refused to continue paying for the goods via the letter of credit and brought action in the courts of Bangladesh against its bank for issuing and approving the letter of credit. In addition, due to the outstanding proceedings, the Bangladeshi customs authorities refused to unload the goods or release the containers from the port without a court order.[37]

While the containers were grounded in Bangladesh, A made several communications to B in order to inquire as to their whereabouts and request their return. The carriage contract between A and B included a standard demurrage clause that provided that B must pay a fixed daily sum for possession of the containers beyond a given fourteen-day 'grace period' at the port of destination. B communicated to A that it was unable to release the containers. A made several inquiries since, including an offer to B to purchase the containers from A once the accumulated demurrage costs exceeded the containers' actual worth. B refused payment and withheld payment of the demurrage costs, thus committing a repudiatory breach.[38] A's continuous requests of payment since were thus interpreted as a rejection of the repudiation.[39]

At trial, Leggatt J held that the 'legitimate interest' qualification applied to quantify the actual amount of debt owed by B to A. In attempting to maintain this conclusion alongside his ruling that the clause is penal,[40] a startling novelty in itself,[41] Leggatt J injected two novel concepts into the discussion. Firstly, Leggatt J opined that:

> the Carrier had a legitimate interest in keeping the contracts of carriage in force for as long as there was a realistic prospect that the Shipper would perform its remaining primary obligations under the contracts by procuring the collection of the goods and the redelivery of the containers. Once it was quite clear, however, that the Shipper was in repudiatory breach of these obligations and that there was no such prospect, the Carrier no longer had any reason to keep the contracts open in the hope of future performance.[42]

[37] ibid at [2]–[10].
[38] ibid at [34], [102].
[39] ibid at [102].
[40] ibid [116]. Note that Leggatt J's review of the law on penalties was in light of the Court of Appeal decision in *Makdessi v Cavendish Square Holdings BV* [2013] EWCA Civ 1539, which has since been overturned by the House of Lords in *Cavendish Square Holdings BV v Makdessi* [2015] UKSC 67.
[41] Jonathan Morgan, '*Smuggling Mitigation into the White & Carter v McGregor: Time to Come Clean?*', (2015) LMCLQ 575, 590.
[42] *MSC Mediterranean* (n 36) [104].

In doing so, Leggatt J essentially suggested a divergent test for the existence of a legitimate interest than what was deployed beforehand. This passage indicates that a legitimate interest would be the existence of a 'realistic prospect' that the repudiating party perform its remaining primary obligations under the contract. Though the operation of this concept of 'realistic prospect' is unapparent from prior appellate court authorities and cannot be seriously considered as a novel alternative formulation of the qualification, it can plausibly be subsumed into the existing definition of 'legitimate interest' without much difficulty if limited to the specific fact patterns in which it is clear when the 'realistic prospect' in performance terminates. More importantly, since Leggatt J held that the demurrage clause was penal, A's decision to keep the contract alive was held not in 'good faith' and thus not a legitimate interest in performance.[43] The invocation of 'good faith' as a corollary to 'legitimate interest' is an unexpected and unnecessary turn, and contributes nothing but more uncertainty to the law on repudiations.

Leggatt J analogised the election of an innocent party facing a repudiation with that of the holder of a discretion under contract. He concluded that the principles requiring 'good faith' and against capricious or arbitrary usage developed to apply to the latter ought apply to the former. His argument is inspired by the decision of the Supreme Court of Canada in *Bhasin v Heynew*.[44] In *Bhashin,* the Supreme Court of Canada held that a minimum standard of honesty is required of contractual parties, which entails '[having] appropriate regard to the legitimate contractual interest of the contracting partner.'[45]

However, this analogy is unsustainable for two main reasons. Firstly, the right of the innocent party to choose whether or not to accept a repudiation does not come about by virtue of any form of agreement, but rather by operation of law in response to a unilateral breach of one. The cases relating to contractual discretion all concern an agreement between the parties to grant the discretion.[46] Therefore, expectations may be a somewhat relevant factor. However, when a contract is repudiated there is no agreed discretion and the innocent party's ability to elect to reject the repudiation is better viewed as a form of remedy offered by the law which gives serious and adequate weight to the binding nature of contracts. This form of discretion is one-sided, where one party has a level of unregulated ability to independently alter the terms of the agreement in a way to which the other party gives its consent in advance. If we analogise scenarios of repudiations to discretion, it is apparent that both parties have the 'discretion' to repudiate.

[43] ibid [97], [98], [118].

[44] [2014] 3 SCR 494.

[45] ibid 499.

[46] *The 'Product Star' (No 2)* [1993] 1 Lloyd's Rep 397*; Paragon Finance Plc v Nash* [2002] 1 WLR 685; *Socimer International Bank Ltd v Standard Bank London Ltd* [2008] 1 Lloyd's Rep 558*; British Telecommunications Plc v Telefónica O2 UK Ltd* [2014] UKSC 42.

However, 'discretion' of whether to keep the contract alive or not is in response to a wrong done by the other party. It is a consequence of the other party's choice to repudiate. In repudiating, the first party triggers the right of the second party to choose whether to accept or reject the repudiation. This right was not exercised by free choice; it is therefore not truly a 'discretion'. Professor Janet O'Sullivan accurately points out that the repudiating party knows its repudiation can either be accepted, requiring it to pay damages, or rejected, and the contract will subsist.[47] In either case, it is awarded a degree of certainty as any liability it may incur will be framed by the contractual agreement while any discretion under a contract can create an inability of the other party to predict the extent of its obligations.

Second, the gist of the dictum in *BT v Telefonica O2*,[48] one of the decisions relied upon by Leggatt J in his analogy, effectively goes against any such analogising. In *BT v Telefonica O2,* Lord Sumption's argument in favour of the requirement of good faith in exercise of contractual discretion was rooted in the view that the parties must act consistently with the contractual purpose.[49] This is fatal to the analogy since repudiation is, in itself, clearly inconsistent with the contractual purpose. Therefore, it would be illogical and unduly onerous to require the innocent party to act consistently with a contractual purpose that has been blatantly disregarded by the repudiating party. Furthermore, in rejecting a repudiation, the innocent party can be seen as acting in a manner consistent with the main underlying purpose of the contract, which is performance.

Leggatt J's resort to the concept of good faith is thus simply untenable. However, it is not entirely incongruous. The usage of terms such as 'wholly unreasonable' or 'perverse' in the case law can be seen as semantic substitutes for 'in *mala fide*'. The consistent affirmation of a purely financial interest as 'legitimate' makes it a herculean task to differentiate between parties who seek performance of the repudiated contract out of genuine financial need and parties that seek performance out of 'malice' or cynically. It seems as though Leggatt J's usage of 'good faith' was necessary in his view so to avoid sterilisation of the qualification. However, 'good faith' is strikingly unmerited if we look to the source of the confusion; that, financially speaking, opting to accept the repudiation and mitigating the losses is the soundest route to take. A purely financial interest is, in fact, not a legitimate interest at all.

4. A Financial Interest Is Not a Legitimate Interest

The murkiness of the 'legitimate interest' terminology results from the courts' consistent adherence to the proposition that a legitimate interest may be a purely financial one. This conflation of a purely financial interest with a 'legitimate

[47] Janet O'Sullivan, '*Keeping the Contract Alive: Unaccepted Repudiation and the Protection of the Performance Interest*' (2016).

[48] *BT v Telefónica O2* (n 46).

[49] ibid [37].

interest' is unsound. The analysis below demonstrates that a purely financial interest is not a 'legitimate' interest.

In *White & Carter*, Lord Reid insinuated that a legitimate interest should be weighed against claiming damages.[50] Under English law, damages are assessed with reference to principles of remoteness, causation, and whether or not the innocent party attempted to mitigate the losses it sustained. Sometimes referred to as the 'duty' to mitigate, it arises upon a termination of the contract due to a breach composed of three propositions.[51] Firstly, the claimant cannot recover for losses it could have taken reasonable steps to avoid. Second, the claimant can recover any losses incurred in taking such reasonable steps to avoid the loss. Third, the claimant cannot recover for losses it has successfully avoided by virtue of those steps.[52] However, once an innocent party rejects a repudiation, the contract subsists and is not terminated by the breach. Therefore, no duty to mitigate losses accrues. Yet, practically speaking, if the 'legitimate interest' is purely financial, the rejecting party would be conspicuously financially better off had it taken the route of mitigation and damages by accepting the repudiation for several reasons.

First, mitigating the loss is the commercially sound thing to do. By mitigating, the innocent party is prevented from solely relying on the uncertain justice system in order to retrieve its money or on the slim chance that the repudiating party will turn around and perform its obligations. Legal proceedings are lengthy, during which time the rejecting party operates with increasing deficit. In *The Alaskan Trader*, for example, the ship-owners lost time in which they could have chartered the ship and incurred substantial legal costs.[53] It is in a commercial party's best interest to limit its financial losses for the sake of its own operations and its other business endeavours. This is also the downfall of Leggatt J's concept of 'realistic prospect' of performance in *MSC Mediterranean*.[54] It is financially unsound to rely on an expectation that a repudiating party will nonetheless perform. It is equally financially unsound to expect a company facing a repudiation to hold an assessment of whether there is a 'realistic prospect' of performance on part of such a party whose conduct or communication indicate that it does not plan to perform at all. The loss potential when mitigation is pursued is always smaller than the loss potential when mitigation is not pursued.[55] Damages are meant to put a party in the position it would have been at had the contract been performed.[56] This would mean that the mitigating party receives money for its endeavours in attempting

[50] *White & Carter* (n 4) 437.
[51] O'Sullivan & Hilliard, *The Law of Contract*, (6th edn OUP 2014) 411–412.
[52] *Hussey v Eels* [1990] 2 QB 277.
[53] *The Alaskan Trader* (n 20).
[54] *MSC Mediterranean* (n 36) [104].
[55] Bridge (1989) 105 LQR 398, 399–410.
[56] *Robinson v Harman* (1848) 1 Exchequer Reports (Welsby, Hurlstone and Gordon) 850, 855.

to mitigate the loss even if unsuccessful.[57] The difference in value, which in the worst case would be full performance, covers what would have been obtained by unilateral performance.

Second, failing to mitigate is especially financially insensible if the repudiating party did so due to lack of financial means to perform the contract. Such a party is likely to go insolvent, which would result in the rejecting party only being able to recover a much reduced amount out of the repudiating company's remaining assets. If the former does not attempt to mitigate, its losses would not only be significantly larger, but also partially, and potentially wholly, unrecoverable.

Third, there is great advantage to be had in quick retrieval of some of the losses sustained. In fact, the value of the availability of money has been recognised in *Sempra Metals v IRC*,[58] a restitution claim, as an 'enrichment' which can be claimed for. Lord Hope of Craighead held that:

> ...the enrichment consists, not of the payment of a sum of money as such, but of its payment prematurely… It was the opportunity to turn the money to account during the period of the enrichment that passed from Sempra to the revenue. This is the benefit which the defendant is presumed to have derived from money in its hands.[59]

This value is equally inherent in instances where a party to a contract faces a repudiation. In *The Alaskan Trader*, the main catalyst to finding no legitimate interest was that the arbitrator so ruled and the court could not find an error in the arbitrator's logic.[60] It remains that there is no error of logic to be found. The fact that, after undertaking repairs at a considerable cost of £800K, the ship-owner then fully staffed the ship and kept it waiting for the repudiating charterers for several months is innocuous to say the least. After sustaining such a financial deficit, a reasonable ship-owner would attempt to try and 'make a quick buck' to start covering for the hefty losses sustained. This would help avoid operating with a large budgetary hole and the availability of money can be used to support an expensive and lengthy process of litigation.

In addition, including 'financial interest' as a legitimate interest undermines the rationale to allowing unilateral repudiations of a contract at all. Repudiation serves an important economical role; it allows the repudiating party to mitigate its future losses arising from the performance of the contract. Professors Oren Bar Gill and Omri Ben-Shahar provided an elegant economic analysis of the

[57] *Gebruder Metelmann GmbH & Co v NBR (London) Ltd* [1984] 1 Lloyd's Rep 614.
[58] *Sempra Metals Ltd. v IRC* [2008] 1 AC 561.
[59] ibid 586.
[60] *The Alaskan Trader* (n 20).

credibility of threats to breach contracts, which in a simplified form holds that if the cost of damages is lower than the cost of performance a party ought breach and therefore any antecedent threats to breach are more credible.[61] This analysis operates smoothly in this area of law as well. If the cost of performance exceeds the amount of damages the company would have to pay, it should repudiate. The ability to repudiate gives expression to economic reality (financial hardships, competition, etc.). Indeed, the charterers in *Alaskan Trader* and *The Aquafaith* sought to mitigate losses by repudiating the charter party contract.[62] Combined with the duty to mitigate on the innocent party, the end result of an accepted repudiation is that it allows both parties to hedge their losses and gains. By including a purely financial interest as a legitimate interest, the courts effectively prevent this positive result by permitting too wide a range of interests to facilitate rejections of repudiations, which are essentially demands for specific performance.[63] The courts have effectively ignored Lord Reid's proposition that the legitimate interest should be weighted against claiming damages, a proposition that would have helped the courts evaluate the legitimacy of financial interests. As in a system of law dependant on how a claim is made, as English common law is, a commercial party can present any interest as 'financial', it is best to do away with considering purely financial interests as legitimate.

If the 'legitimacy' of the interest is measured according to the commercial sensibility of the decision to reject the repudiation, it is clear that a purely financial interest lacks entirely in requisite legitimacy. It is simply illogical for a commercial party to risk heftier losses for no more potential gain. A financial interest cannot, therefore, be regarded as a 'legitimate' interest, as, in light of the above, it is wholly financially unsound for the party to not to make any attempt to mitigate or alleviate its losses. Professor Morgan views the principle of mitigation as supporting the abolition of *White & Carter* altogether.[64] He strongly advocates that the mitigation principle presides over contractual remedies in a way that does not allow the innocent party to reject a repudiation, as it is thus caused a loss 'it could have easily avoided'. Morgan holds that a requirement to mitigate should arise automatically upon a repudiation,[65] and that too much emphasis is recently placed on the value of performance.[66] Nevertheless, there is no need to amputate the leg of a patient with merely a sore toe. There are many instances where it is unrealistic to claim

[61] Omri Ben-Shahar & Oren Bar-Gill, 'Threatening an Irrational Breach of Contract' (2003) 11 Supreme Court Economic Review 143.

[62] *The Alaskan Trader* (n 20); *The Aquafaith* (n 31).

[63] *The Puerto Buitrago* (n 24) 259.

[64] Jonathan Morgan, 'Smuggling Mitigation into the *White & Carter v McGregor*: Time to Come Clean?' (2015) LMCLQ 575.

[65] ibid 584, 590.

[66] ibid 583.

that proper adequate mitigation can be expected of an innocent party. In such cases, a 'Writ in water' rejection should be made available precisely because the nature of the innocent party's interest in the contract indicates that there is no feasible way to mitigate for its loss; for example, where that party is an employee or holiday goer. For this reason, a 'legitimate interest' qualification that excludes purely financial interests will be useful to differentiate those worthy parties from those who should be subjected to a duty to mitigate outright.

5. What Should Constitute a 'Legitimate Interest'

By ruling out financial interest as a legitimate interest, one admittedly kicks a hornet's nest by insinuating that repudiation ought be accepted other than in very specific fact patterns. In wholly commercial endeavours, such an assumption is sound in light of the points presented above. When the contract is not entirely for financial gain but provides for additional gains of a different kind, the 'legitimate interest' comes into play. It is where the principal value of performance is one for which damages cannot easily account and mitigation cannot be equally successful in achieving that a legitimate interest should be found. If a financial interest is rejected as a legitimate interest, one must evaluate what *types* of interests could qualify as legitimate. In *White & Carter*, Lord Reid gave one example of a report being prepared by an expert after he is informed that it is no longer necessary, but it was primarily to present the paradigm case of an obligation to pay that is dependent on performance.[67] Other than that, it is evident from the case law that, in attempting to define a legitimate interest, there has been little attempt to consider the specific interests that might answer that test.

Watts v Morrow provides an interesting, comparative, starting point.[68] In *Watts* the claimants attempted to mount a claim for damages for the distress they suffered as a result of breach of a survey contract. Bingham LJ held that there were two exceptions to the general rule that non-pecuniary losses cannot be recovered,[69] the first of which is of particular relevance here.

Bingham LJ held that non-pecuniary interests could be recovered where 'the very object of a contract is to provide pleasure, relaxation, peace of mind or freedom from molestation.'[70] This exception represents an acknowledgement of values to performance other than financial gains that may be held by a party to a contract. Such a 'value' should represent the legitimacy of an interest to reject a

[67] *White & Carter* (n 4) 428.
[68] [1991] 1 WLR 1421.
[69] ibid 1445.
[70] ibid.

repudiation—a principal value to performance for which damages cannot easily account and mitigation cannot be equally successful in achieving.

Many decisions and doctrines are based on different 'values' provided for by unique contracts,[71] including freedom from molestation,[72] secrecy of information,[73] and pleasure and relaxation.[74] In *Jarvis v Swan Tours*, the disappointment in not obtaining the benefit of enjoyment was accounted for by the Court of Appeal.[75] It would probably have been accounted for had Swan Tours simply repudiated the travel contract and not merely poorly performed since the court recognised that the breach was not fundamental and nevertheless opted to grant damages for distress and disappointment.[76] Similarly, in cases where one company concludes an advertising contract with another and the other repudiates, the company seeking exposure has a legitimate interest in rejecting the repudiation. The value of exposure at the desired period of time would be an interest not fully accountable in damages. The loss of exposure time at a period when such exposure was critical for the company is neither included in the costs of finding a new advertising contract nor can it be effectively mitigated. Advertising benefits are not impossible to assess at a particular point in time but their cumulative effect in obtaining new customers, maintaining existing ones, and creating a 'loyalty' culture is likely to be considered too remote for a damages award for a single repudiated advertising contract.

The main criticism of such a formulation is that it narrows down the election. However, this formulation of 'legitimate interest' is not as narrow as may seem. A court may find that a 'legitimate interest' in rejecting repudiation is commercial reputation and the desire not to seem commercially unreliable or unstable. Such a value may be considered such that is not properly ascertainable by way of damages. Accepting commercial reputation as a legitimate interest will widen this formulation to encompass cases where the rejection of the repudiation is due to a calculated assessment and help set apart the malicious from the genuine performance seekers without the unnatural resort to 'good faith'. The initial teething problems of such a formulation of 'legitimate interest' are evident, especially in the context of employment contracts, but its result would be a much higher degree of certainty for contracting parties. Claimants will be aware of the circumstances in which they hold a 'legitimate interest' allowing them to reject a repudiation. The category should not, however, be closed. The courts should allow for introduction of new 'legitimate interests' as new fact patterns emerge in order for this formulation of the qualification to have the most positive effect.

With this approach, a difference may emerge between judicial treatment of elections to accept or reject a repudiation and elections to accept or reject

[71] *Farley v Skinner* [2002] 2 AC 732, 753.
[72] *Watts v Morrows* (n 68).
[73] *Attorney General v Blake* [2000] 4 All ER 385.
[74] *Jackson v Horizon Holidays* [1975] 1 WLR 1468.
[75] [1973] QB 233.
[76] [1973] QB 233, 240 per Stephenson LJ.

repudiatory breaches. The categories of 'legitimate interests' for the former is likely to be more restrictive than that of the latter. For example, in the employment context, an employee rejecting a repudiatory breach by an employer could have the 'legitimate interest' of 'having an occupation' or even dignity, while these would not be accepted as a 'legitimate interest' in instances of fully fledged dismissals. This possibility of a more nuanced approach is appropriate when considering the variety of possible interests in many different sectors that the courts will face. This flexibility brought about by the suggested formulation of 'legitimate interest' will not create uncertainty since there will be a clear yardstick for legitimacy. The variety of interests will all have to represent a principal value to performance that is not compensable by way of damages or can be easily mitigated. An approach that is both nuanced and flexible, but also creates certainty at the same time is a desirable outcome. Such an outcome is not achieved by the 'reasonableness' alternative.

6. Considering the 'Reasonableness' Alternative

Although the language of reasonableness has been often utilised in the case law,[77] the reasonableness concept does not succeed in achieving the same outcome as the reformulated 'legitimate interest'. A 'reasonableness' standard looks into the defendant's behaviour in rejecting the repudiation and it is usually measured against peers such as 'the reasonable man'[78] or 'the officious bystander'.[79] Standards of behaviour are measured in a comparative fashion, and seek to find the average conduct of, in this instance, commercially sound businessmen. However, in the case of rejected repudiations it is much more precise to look into the reason for the rejection itself, rather than as to what a 'reasonable businessman' do in response to the circumstances. It is less able to result in a clear delineation between circumstances of a rejection that enable keeping the contract alive and those which do not.

Further, if we maintain that the election of the innocent party whether to accept or reject the repudiation is a free one, it should be able to make it even if the reasonable man would not have. The main question is not whether the choice itself was reasonable, but rather whether the party had an interest legitimate enough to entitle it to make that choice.

Finally, unlike other areas of contract the law, where these tests are used to resolve ambiguities in party intention, here a party makes a clear-cut decision. There is therefore no need to speculate what that party would have done; this will lead into an examination as to *why* the party acted the way it did and whether its reasons are meritorious. It will provide for a redundant step before naturally proceeding to examine the interests of the rejecting party and will eventually require deployment

[77] *The Puerto Buitrago* (n 24); *The Odenfeld* (n 19).

[78] *ICS v West Bromwich Building Society* [1998] 1 W.L.R. 896, 913 per Lord Hoffman.

[79] *Shirlaw v Southern Foundries* (1926) Ltd [1939] 2 KB 206, 227–8 per MacKinnon LJ.

of a concept akin to 'legitimate interest'. A straightforward examination into the legitimacy of the interest can come up with answers irrespective of sometimes artificial ponderings as to how many or how few reasonable men would have made the same choice.

7. Conclusion

The main issue with the 'legitimate interest' qualification to the innocent party's ability to elect whether to accept or reject a repudiation is that it requires clarification. Since the decision in *White & Carter*,[80] the language of the courts in applying it has been unnecessarily ambiguous and confusing. Neither the current formulation of 'legitimate interest' post-*Aquafaith*,[81] nor Leggatt J's dictum in *MSC Mediterranean v Cottonex*[82] are loyal to Lord Reid's original intention. Furthermore, both are unsatisfactory. The *Aquafaith* and the preceding line of cases are unsatisfactory due their misguided use of 'reasonableness' terminology to define a 'legitimate interest' and their recognition of a purely finanial interest as a 'legitimate interest'. The *MSC Mediterranean* decision is similarly unsatisfactory due to its unsuccessful attempt at solving the difficulties brought about by the recognition of a purely financial interest as a 'legitimate interest' through an untenable analogy to 'good faith' and the unnecessary concept of 'realistic prospect of performance.'

This lack of clarity is caused by the express recognition of purely financial interests as legitimate. Lord Reid's dictum conveys that the legitimacy of the rejection should be weighted against claiming damages.[83] While this was mentioned by Lord Denning in *The Puerto Buitrago*[84] and applied by the arbitrator in *The Alaskan Trader*,[85] it has since been disregarded by the courts. Instead, the courts opted to hold that *any* financial interest is valid irrespective of whether the rejecting party would have been equally better off by mitigating and claiming damages.

A purely financial interest must not be considered a legitimate interest. It is financially unsound for a company not to attempt any form of mitigation and opt for the uncertain route of litigation. That is especially true when the repudiating party does so due to financial difficulties. A rejection of a repudiation for solely a financial interest is a risky bet which can result in no more than the financial value of performance, a value that the much safer and more certain route of mitigation and damages provides. The ability to claim a financial interest as legitimate without any comparison to damages means that a rejecting commercial party can frame its claim so that its interest is always legitimate. This is evident by the case law and the rarity of a finding that an interest is illegitimate. The recognition of a

[80] *White & Carter* (n 4).
[81] *The Aquafaith* (n 31).
[82] *MSC Mediterranean* (n 36).
[83] *White & Carter* (n 4) 431.
[84] *The Puerto Buitrago* (n 24).
[85] *The Alaskan Trader* (n 20).

purely financial interest as legitimate results in the unfortunate serialisation of a potentially highly valuable and useful qualification. What constitutes a 'legitimate interest' should therefore be redefined.

This redefinition should not take the form of replacing the qualification altogether with a 'reasonableness' test. While it is capable of testing whether an interest is 'legitimate', it will not produce a sufficient level of certainty. Moreover, a 'reasonableness' test is redundant as it will nevertheless require deployment of an analysis of the legitimacy of the interests in place. Overall, the test in instances of rejected repudiations does not address an ambiguity in party intention. The test must be more precisely described as appertaining to the rationale behind the choice made, rather than to whether it would have been the popular choice amongst peers.

Defining a 'legitimate interest' as requiring a principal value to performance for which damages cannot easily account and mitigation cannot be equally successful in achieving is the soundest solution. The principal value can take many forms and does not overburden the field by placing strenuous limitations on the validity of rejections of repudiations. On the contrary, it affords the courts requisite flexibility to respond to novel fact patterns along with the certainty of a clear yardstick. If the party's interest in performance has been recognised in previous cases as compensable by way of damages and it can be expected to take reasonable steps to mitigate for it, the interest is illegitimate. Maintaining this alternative formulation is preferable to imposing a duty to mitigate as it responds better to situations in which it is impracticable to require the innocent party to mitigate. Delineating the types of interests that are legitimate will result in law that is, on the one hand, more nuanced, and, on the other, more certain.

Rules of Engagement and the Use of Force in United Nations Operations

John Simpson[1]

1. Introduction

THE MANDATE ISSUED by the Security Council ('SC') establishes the legal basis for the use of force in a United Nations ('UN') mission, while the Operational Plan and Rules of Engagement ('ROE') are the instruments used to implement that mandate. Consequently, the ROE of any mission cannot exceed the purview of the mandate given to that mission by the SC.[2] ROE are accepted by contributing nations and the UN as the most common and effective way to control the use of force by military forces during UN operations.

The ROE set out when and how the use of force is authorised. They also reflect the unique capability of available weapons systems and the specific rules of international customary and treaty law that may be applicable. Whereas the mandate reflects the political, diplomatic, policy, and operational objectives of the mission, the ROE represent the practical application of those mandates where the use of force has been chosen as a means of implementation. However, when the mandate assigned by the SC does not match the ROE required by military units on the ground, then it becomes more difficult to carry out the intent and will of the UN and the SC.

The gap between mandate creation by the SC and the application of ROE by individual contributing nations in compliance therewith is a fatal flaw of the mandate system. When the ROE required to protect human rights and international humanitarian law in an area of conflict—a fundamental purpose of

[1] B.C.L./LL.B, McGill University. John.simpson2@mail.mcgill.ca. I would like to thank Professor Payam Akhavan for his guidance and support.

[2] Terry Gill, 'Legal Parameters for the Use of Force in the Context of the UN Collective Security System' in Terry Gill and Dieter Fleck (eds), *The Handbook of the International Law of Military Operations*, (Oxford: Oxford University Press 2010) fn 5 112.

the UN—do not match the mandate given by the SC, problems arise.[3] A mission may freeze due to its inability to act, or it may undergo 'Mandate Creep' where a Chapter VI mandate, meant to focus on observation and civil support, will begin to shift into Chapter VII, which allows military intervention, without the approval of the SC. In the former case, UN soldiers are forced to stand idly by and watch human rights abuses without lifting a finger, while in the latter, UN forces arguably violate international law despite furthering the true intent and purpose of their mission. The UN's experiences in Rwanda and the former Yugoslavia attest to these unfortunate situations. The UN Organization Mission in the Democratic Republic of the Congo (MONUC) suffered from the same deficiencies, but made decisive changes after the intervention by French forces in 2003. This particular mission will be elaborated upon more extensively below.

Regardless, the disparity between mandates and ROE development reflects a need for change. This article will discuss the interrelationship between UN mandates and ROE in situations where the UN has deemed the use of force to be necessary. Section 2 will discuss basic principles of ROE and how they are formulated. Section 3 will then analyse how the use of force and ROE fit into UN mandates as well as international humanitarian law ('IHL'). Section 4 focuses on the political and policy influences that negatively affect the creation of mandates and how those influences create gaps between the conceptualisation of the mandate and the formulation of ROE for UN missions. Section 5 explores the consequences of those gaps: the phenomena of Frozen Mandate and Mandate Creep. Section 6 will provide a case study, based on the MONUC mission, which concludes that clear mandate formulation and standardisation at the outset of UN missions, rather than progressive escalation of mandates, can have positive and tangible results for ROE drafting and implementation.

2. Rules of Engagement: Basic Principles

ROE constitute the most direct influence that the international law of armed conflict has on peacekeepers.[4] In the context of UN peacekeeping missions, the ROE are inherently connected to the level of force authorised in the mandate issued by the Security Council, which is, in turn, influenced by the concerns of the Security Council.[5]

[3] *Charter of the United Nations* (adopted 26 June 1945, entered into force 24 October 1945) 1 UNTS XVI article 1(1).

[4] Mary Ellen O'Connell, 'Historical Development and Legal Basis: Binding Effect of International Law for the Soldier' in Dieter Fleck (ed), *The Handbook of International Humanitarian Law* (Oxford: Oxford University Press, 2013) 38.

[5] The term peacekeeping is used in this paper as a term encompassing peacekeeping in its classical sense as well as the principle of peace enforcement.

The application and translation of ROE is the responsibility of the commander and their planning staff and, especially in the UN context, must be translated from a theoretical mandate into a practical, understandable, and applicable set of rules on the permitted use of force.[6] Commanders are obliged to understand and apply all applicable rules of international law and formulate ROE that comply with them. They must also ensure that ROE given to soldiers in the field are understandable so that, if soldiers do have to use force in the course of their mission, it is done in a way that is both controlled and legal.[7] The United States Judge Advocate General's *Operational Law Handbook* comments on the multifaceted influences that must be taken into account during the establishment of ROE, including customary and treaty law principles, but also political objectives and mission limitations such as—in the context of a UN mission—the applicable Security Council mandate.[8] The ROE stipulate the means and methods of permitted use of force and are normally formulated by the Department of Peacekeeping Operations, with input from the Force Commander, and issued to troops in the field in the form of a pamphlet or laminated card known as Orders for Opening Fire.[9]

The applicable rules for using force under the ROE will differ according to international law and according to each new mission, as well as each individual situation that occurs, but in all cases ROE and the application of force must respect the IHL principles of necessity, distinction, proportionality, and humanity, topics that will be discussed in more detail in Section 4 of this article.[10] Therefore, the soldiers tasked with implementing the ROE must be familiar with, and trained in, the ROE of their particular mission. ROE for UN missions are often defined in negative terms, as rules of when not to use of force. This reflects the statement of Ben F. Klappe, in his section of *The Handbook of International Humanitarian Law* on ROE, expressing the general hesitancy of soldiers in UN missions to use force in any proactive way, even if doing so would serve the mission's purpose and the greater good more fully. He found that soldiers taking part in UN missions were trained to be reluctant to use force at all.[11]

This illustrates the conflict between mandates and ROE that has led soldiers to follow their ROE too restrictively or act on their principles of humanity and

[6] Patrick Cammaert and Ben Klappe, 'Application of Force and Rules of Engagement in Peace Operations' in Gill and Fleck (eds) (n 2) 151; see, also, Judge Advocate General of the United States, *United States Operational Law Handbook* (Virginia: The Judge Advocate General's Legal Center and School, U.S. Army 2013) 75.

[7] Ben F. Klappe, 'The Law of International Peace Operations: Rules of Engagement' in Fleck (ed) (n 4) 634.

[8] Judge Advocate General of the United States, (n 6) 75.

[9] Trevor Findlay, 'The Use of Force by Peacekeepers in Self-Defence: Some Politico-Legal Implications' in Alex Morrison et al (eds), *Peacekeeping with Muscle: The Use of Force in International Conflict Resolution* (Clementsport: The Canadian Peacekeeping Press 1997) 52.

[10] Alan Cole et al (eds), *Rules of Engagement Handbook* (Sanremo: International Institute of Humanitarian Law 2009) 5.

[11] Klappe, in Fleck (ed) (n 5) 633.

conscience, resulting in a distortion of their ROE as well as the mandate of the mission.[12] The use of force is ultimately the soldier's decision, and ROE must be permitted the flexibility to adapt to the specificity of each mission and the obstacles encountered. As will be seen in the following Sections, ROE for UN missions suffer from being limited by the mandate that gives them legal legitimacy.

3. The Use of Force in UN Missions and International Law

ROE represent the practical application of the use of force theoretically envisioned by the United Nations and are therefore influenced by political, legal and social factors.[13] While UN troops are, in all situations, permitted to use force up to and including lethal force for their own personal self-defence as well as the defence of their unit, beyond those parameters it is the mandate that dictates how far individual soldiers are permitted to use force to defend civilians, take proactive roles, or—in a more recent and controversial development—defend the mandate and the mission itself.[14] In the UN's *Handbook on United Nations Multidimensional Peacekeeping Operations* the 'appropriate use of force' is explained in the following way:

> The use of force by the military component will depend on the mandate of the peacekeeping operation and the rules of engagement; sometimes the Security Council will authorise a peacekeeping operation to use armed force in situations other than in self-defence. The circumstances under which the operation may use armed force will then be spelt out in the relevant resolution of the Council. The rules of engagement for the peacekeeping operation will clarify the different levels of force that can be used in various circumstances, how each level of force should be used and any authorisations that may need to be obtained from commanders.[15]

In all situations where UN troops are permitted to use force, whether in self-defence or in a peace-enforcement capacity, the fundamental principles of IHL will apply.[16] Peacekeeping falls under a Chapter VI mandate, generally a role of observation and civilian support, or a Chapter VII mandate, which normally calls for more aggressive action in conflict zones where human rights abuses are occurring.

[12] Mary Ellen O'Connell, in Fleck (ed) (n 5) 39.

[13] Chief of the Defence Staff, *Use of Force for CF Operations, B-GJ-005-501/FP-001* (Ottawa: Government of Canada 2008) 2–4; see, also, Judge Advocate General of the United States (n 6) 75; see, also, Cammaert and Klappe, in Gill and Fleck (eds) (n 2) 151.

[14] Trevor Findlay, *The Use of Force in UN Peace Operations* (Oxford: Oxford University Press 2002) 361.

[15] United Nations, *Handbook on United Nations Multidimensional Peacekeeping Operations* (New York: Department of Peacekeeping Operations 2003) 57.

[16] UN Doc ST/SGB/1999/13 (1999), *Secretary-General's Bulletin: Observance by United Nations Forces of International Humanitarian Law* ('*Secretary General's Bulletin*').

A. Chapter VI

The use of force for self-defence is considered the minimum that the UN can afford to the troops it sends into conflict zones.[17] This basic principle was established during the deployment of the UN Emergency Force (UNEF I) that was sent to Egypt in November 1956, and has been included in the guidelines of every peacekeeping mission deployed by the UN since. This mission established the classical conception of peacekeeping, based on Chapter VI of the UN Charter, the idea for which is largely credited to then Canadian Minister of Foreign Affairs Lester B. Pearson. Chapter VI of the UN Charter allows the SC to use any means, short of calling for the use of force, to peacefully resolve a conflict.[18] The self-defence principle has, however, been blurred and expanded since that first deployment to include defence of the mandate.

The principle of self-defence in the context of UN Chapter VI mandates has expanded to include, depending on the capacity of the UN force, defence of the mandate and the mission.[19] The expansion of the definition led former UN Secretary-General (SG) Dag Hammarskjold to refer to peacekeeping as 'Chapter VI-and-a-half'.[20] The extension of the self-defence principle to defence of the mission and the mandate gives UN forces an additional guiding principle to help them ascertain hostile intent. Allowing these principles to apply to the defence of the mandate is a positive development, as it gives UN forces the interpretive flexibility to act where the substance and purpose of a mission is threatened.[21] The principle of defence of the mission was used during the UN Operation in the Congo (ONUC) in the 1960s and was criticised as being an unjustified expansion of a Chapter VI mandate into Chapter VII.[22] It has developed since then as the basis for the new, proactive role that UN peacekeeping forces increasingly have to play; this will be analysed in detail through a case study in Section 7. Unfortunately, it comes nowhere close to bridging the gap between the mandate creation process and ROE formulation. UN peacekeepers are trained to be reluctant to use force, and this, combined with the predominantly defensive nature of UN ROE, creates hesitancy in using force even when it is justified. For example, during ONUC twelve Irish soldiers were ambushed and ten were killed. In the field report, the Irish officer's confusion over their ROE was cited as the reason for the incident;

[17] Findlay, in Alex Morrison et al (eds) (n 9) 53.
[18] Ben F. Klappe, 'The Law of International Peace Operations: General' in Fleck (ed) (n 4) 612.
[19] Terry Gill, 'Legal Parameters for the Use of Force within the Context of Peace Operations' in Gill and Fleck (eds) (n 2) 150.
[20] Larry Johnson, 'Peacekeeping with Muscle: Possibilities Under the Charter of the United Nations' in Morrison et al (eds) (n 9) 27.
[21] Findlay (n 17) 87.
[22] ibid 88.

the troops felt that they only had legal authority to use force if fired upon first.[23] During the UN Assistance Mission for Rwanda (UNAMIR), in addition to the almost 800,000 civilians massacred, ten Belgian soldiers, who were part of the UN force, were disarmed and executed by Hutu militiamen during the Tutsi genocide without firing a shot in self-defence.[24] Since the Hutu militiamen did not use force in apprehending the Belgians, the latter did not feel that they could use force legally according to their mandate and, as a result, handed over their weapons.[25] ROE cannot be so stringent that soldiers second-guess themselves when interpreting whether a potential assailant can be considered a military target. This must be the soldier's decision based on the reasonable belief in the hostile intent of the potential assailant and the particular situation and context.[26] This will likely result in UN forces gaining credibility and effectiveness rather than losing it.

It has always been assumed that the credibility of a UN mission under Chapter VI has derived from the consent of the parties involved.[27] However, this is becoming less and less applicable, as parties to conflicts manipulate the UN's efforts for strategic purposes and undermine the efforts of peacekeepers.[28] In the *Brahimi Report*, chaired by Lakhdar Brahimi, the panel tasked with drafting the report reiterated the importance of consent and impartiality towards the parties in conflict as the primary foundations of peacekeeping operations. This was, however, a precursor to the more important point that consent, or the lack of it, should not be permitted to hinder a UN force's ROE or their application in self-defence, defence of the mandate, or defence of civilians.[29] Brahimi states in his report that 'rules of engagement should not limit contingents to stroke-for-stroke responses but should allow ripostes sufficient to silence a source of deadly fire that is directed at United Nations troops or at the people they are charged to protect and…should not force United Nations contingents to cede the initiative to their attackers.'[30] The *Brahimi Report* argues that UN ROE should allow for proactive engagement with forces that show hostile intent.

[23] For a detailed account of the incident see E. W. Lefever and W. Joshua, *United Nations Peacekeeping in the Congo, 1960-1964: An Analysis of Political, Executive and Military Control*, vol 3 (Brookings Institution for the US Arms Control and Disarmament Agency: Washington, DC 1966) appendix P-7.

[24] Findlay (n 17) 280.

[25] Astri Suhrke, 'Dilemmas of Protection: The Log of the Kigali Battalion' (1998) 5 *International Peacekeeping* 1, 9.

[26] Klappe, in Fleck (ed) (n 4) 635.

[27] Terry Gill, 'Characterisation and Legal Basis for Peace Operations' in Gill and Fleck (eds) (n 2) 136.

[28] UN Doc A/55/305–S/2000/89 (2000), *Identical letters dated 21 August 2000 from the Secretary-General to the President of the General Assembly and the President of the Security Council: report of the Panel on United Nations Peace Operations* ('*Brahimi Report*') para 48; see, for example, how during the UNAMIR mission in Rwanda, the Hutu militia used the pretext of the Arusha Peace Agreement and the limited mandate and ROE of the UN forces to stockpile weapons and prepare for the impending genocide perpetrated against the Tutsi population: Findlay, (n 17) 278.

[29] *Brahimi Report* (n 28) paras 49–50.

[30] *Brahimi Report* (n 28) para 49.

The foundation of UN peacekeeping operations' credibility was previously based on impartiality in the sense of treating both sides equally.[31] The *Brahimi Report* suggests an interpretation of impartiality based on adherence to the principles of the UN Charter and the mandate assigned to the mission by the SC.[32] Where there is an aggressor showing hostile intent, peacekeepers should, under a Chapter VI mandate and their ROE, be able to use force to defend the targets of that aggression with moral and legal justification.[33] This new interpretation of impartiality has the power to lend more credibility to UN missions, as long as mandates remain flexible enough to allow ROE to adapt to changing situations on the ground. It means a UN force with credibility based on positive action instead of a symbolic and non-threatening military presence. In the past, the old view of impartiality pushed some UN commanders to manipulate situations to allow them to intervene where their consciences necessitated action, but the limits of their mandate—and consequently their ROE—prevented the legal use of force.

During the ONUC mission, which took place from 1960-1964, Dag Hammarskjöld used the strategy of interposing UN soldiers between civilians and hostile forces to enable the use of force in self-defence.[34] This way, peacekeepers could use force legally under a Chapter VI mandate if attacked, which helped deter belligerent forces. Hammarskjöld was criticised for allowing this mandate creep in the Congo, which France, Britain, and Belgium claimed was beyond the legal authority of the mission.[35] The UN Protection Force (UNPROFOR) also used this strategy in the former Yugoslavia while it was still mandated under Chapter VI, as did UNAMIR.[36] In Rwanda, Lieutenant-General Roméo Dallaire, the UN Force Commander, posted his small force as guards outside of the safe zone, placing them between fleeing Tutsis and Hutu militiamen.[37] While this strategy shows that UN leaders can, to a certain extent, defend civilians without having to use force, it is clear, based on the frequency of its use as a strategy, that it has become a standard practice, which is cause for alarm.[38] The self-defence principle is not meant to allow commanders to deliberately place UN troops in harm's way

[31] *Brahimi Report* (n 28) para 50.

[32] Ray Crabbe, 'Future Peace Operations: A Conceptual Approach' in Richard Wiggers and Ann L. Griffiths (eds), *Canada and International Humanitarian Law: Peacekeeping and War Crimes in the Modern Era* (Halifax: Centre for Foreign Policy Studies 2002) 111.

[33] *Brahimi Report* (n 28) para 50.

[34] Operations Directive No 8 [untitled], February 1961, UN Archives DAG/13/1.6.5.0.0; see, also, Findlay (n 17) 61.

[35] Findlay, in Morrison et al (eds) (n 9) 65.

[36] Roméo Dallaire, *Shake Hands with the Devil: The Failure of Humanity in Rwanda* (Toronto: Random House Publishers 2004) 268–269; see, also, Victoria Holt and Tobias C. Berkman, *The Impossible Mandate? Military Preparedness, the Responsibility to Protect, and Modern Peace Operations* (Washington: The Henry L. Stimson Center 2006) 83; see, also, D. Last, *Theory, Doctrine and Practice of Conflict De-Escalation in Peacekeeping Operations* (Canadian Peacekeeping Press: Clementsport 1997) 105–107.

[37] Holt and Berkman (n 40) 83.

[38] ibid.

so as to allow reprisal. Using the self-defence principle in such a way glosses over a clearly interventionist act that resembles peacekeeping under a Chapter VII mandate. This phenomenon of 'mandate creep', where a Chapter VI mission is expanded beyond the legal scope of the mandate, will be discussed in Section 5. Whether these commanders openly defied the mandates and ROE that they were given or simply interpreted them as allowing their actions does not matter; what is important is that the commanders were forced into situations where the only way to protect civilians was to bend and distort the mandate of their mission.

B. Chapter VII

Unlike Chapter VI, missions mandated under Chapter VII of the UN Charter do not require the consent of the parties involved and are normally reserved for the most serious situations in which 'international peace and security' are threatened.[39] Originally, for the SC to make such a finding, it required the use of armed force by a state, which the SC deemed to be an act of aggression or a breach of the peace.[40] However, SC policy has progressed since then to recognise that nowhere in Chapter VII does it refer to states as the only entities that can breach international peace and security.[41] Importantly, in the *Report of the International Commission on Intervention and State Sovereignty* ('Responsibility to Protect'), serious threats to civilians and human rights were cited as legitimate reasons for finding a threat to international peace and security, a principle strongly supported by the General Assembly and cited as the grounds for the intervention in Libya in 2011.[42] However, when the SC does mandate a mission under Chapter VII, it rarely expresses how it envisions the enforcement of the mandate, or what levels of force are authorised. The key phrase that denotes a Chapter VII authorisation has become 'all necessary means.'[43]

Even under Chapter VII, the ROE limit the use of force beyond what is necessary to achieve the specific goals mandated in the SC resolution. This means that if protection of civilian populations under immediate threat of violence is not mandated, and thus does not become part of the ROE of the mission, then UN forces will not be legally permitted to defend civilians.[44] For example, in 1993 the SC authorised a Chapter VII mandate for UNPROFOR in the former Yugoslavia, but notably refrained from using the 'all necessary means' language in the resolution authorizing the change in mandate.[45] The resulting application of the mandate

[39] *Charter of the United Nations* (n 3) article 39.

[40] *Certain expenses of the United Nations*, Advisory Opinion, [1962] ICJ Rep 177.

[41] Findlay (n 17) 55.

[42] Gareth Evans et al (eds), *Report of the International Commission on Intervention and State Sovereignty: The Responsibility to Protect* (Ottawa: International Development Research Centre 2001); see, also, UNSC RES 1973 (2011).

[43] UNSC RES 836 (1993); see, also, UNSC RES 1565 (2004).

[44] Findlay (n 17) 425; see, also, Gill, in Gill and Fleck (eds) (n 2) 111.

[45] UNSC RES 807 (1993).

through UNPROFOR's ROE was haphazard and lacked standardisation. Some commanders interpreted their ROE narrowly as being to deter attacks against the safe-areas, but not to defend them where such deterrence failed.[46]

Such a case occurred in the context of the massacres at Srebrenica in 1995. The SC had declared Srebrenica a safe-area and had mandated a contingent of Dutch troops to protect the enclave, and the civilians within it, from attack by Bosnian Serbs.[47] The Netherlands was later pursued in its domestic courts for failing to fulfil their international obligations as peacekeepers. Evidence showed that elements of the Dutch Battalion (Dutchbat) had witnessed more than one instance of Bosnian Serbs beating, and in some instances, killing male refugees outside the compound without taking any action.[48] Therefore, the Dutchbat knew that the refugees were at serious risk of mistreatment should the Bosnian Serbs take control of the camp. Despite this, when the Bosnian Serb Army surrounded the compound, the Dutchbat forces abandoned their positions and withdrew to a nearby compound; in the ensuing massacre, 8,000-10,000 Bosnian Muslim refugees were killed by the Bosnian Serbs forces[49] Interestingly, in a related decision, the Hague Court of Appeal left open the possibility for the UN to be held responsible for failing to fulfil its mandate where the peacekeepers in question are under its effective control. This ruling makes the clarity and enforcement of ROE all the more important in UN missions where the use of force is authorised.[50] It is clear that the formulation of mandates by the SC and their subsequent translation into ROE are in dire need of reinterpretation, both in the case of Chapter VI and Chapter VII mandates.

4. The Relationship Between UN Mandates and Rules of Engagement

The ROE represent the essence of applicable IHL and the SC mandate and, therefore, reveal—more than any other UN document—the true nature of the mission.[51] Unfortunately, one has only to examine the divide between the formulation of the mandate and the creation of the ROE for UN peacekeeping missions to understand why, so often, ROE have proven to be incommensurate with the SC's vision. The actions of a State's peacekeepers have the power to cast the governments that authorised their involvement in a negative light. In Part A of this Section, the result of these policy considerations on the mandate development

[46] Findlay (n 17) 229.

[47] UNSC RES 819 (1993); see, also, UNSC RES 824 (1993); see, also, UNSC RES 836 (1993).

[48] *Nuhanovic v State of the Netherlands*; *Mustafic v State of the Netherlands*, [2011] Court of Appeal of The Hague, Judgment of 5 July 2011.

[49] Klappe, in Fleck (ed) (n 5) 622–624.

[50] *The Association of Citizens Mothers of Srebrenica et al. v The State of the Netherlands*, [2010] Court of Appeal of The Hague, Judgment of 30 March 2010; see, also, Fleck, 'Status of Forces in Enforcement and Peace Enforcement Operations' in Gill and Fleck (eds) (n 2)108.

[51] Findlay (n 17) 369.

process will be analysed, and Part B will focus on how the mandate and IHL, in turn, affect the development of ROE.

A. The Influences that Shape the Creation of the Mandate

Political, policy and diplomatic forces shape the mandate creation process at the UN and have crippling effects for both UN missions and the individuals they are deployed to protect. As will be seen in Part B of this Section, the SC mandate is the main influence on the formulation of the mission's ROE, and, therefore, the influences that shape the SC mandate have a direct impact on the UN force's ability to act on the ground. These hesitations and preoccupations with national image and popular policy will be analysed in the context of the mandates for the UNPROFOR, UNAMIR and UNOSOM II missions in the following pages.

Based on reports published by both the UN as well as independent parties, the international preference for a 'light', largely symbolic peacekeeping presence that dominated the 1990's had a strong effect on the creation and maintenance of both UNPROFOR and UNAMIR.[52] During the collapse of the Dutchbat in Srebrenica and the international community's inaction during the Rwandan genocide, the fresh memory of US engagement in the Battle of Mogadishu during the UN Operation in Somalia II (UNOSOM II) in October 1993, only intensified the pressure to support policies of symbolic as opposed to effective peacekeeping.[53] The result, as was seen in Rwanda, has been the authorisation of small under-armed peacekeeping forces with insufficient legal grounds to use force when needed. In its comprehensive report following the UNAMIR mission, the UN Department of Peacekeeping Operations explained that 'the mandates of UNAMIR were a product of the international political environment in which they were formulated, and tended to reflect concerns and imperatives of certain Member States that had little to do with the situation in Rwanda.'[54] The political tension in the creation of UNAMIR was so palpable that as soon as the Arusha Peace Accord broke down with the assassination of President Habyarimana, France and Belgium sent forces into the country to unilaterally—and without warning—evacuate its peacekeepers and foreign nationals, thus endangering the mission even further.[55]

Political concerns and national interests must always play a role in the decision making process of a sovereign nation, but these political and policy influences must not be permitted to grow to such levels that they hinder the UN's legal ability to

[52] *Brahimi Report* (n 31) paras 19–22.

[53] On October 3, 1993, soldiers of the United States Rangers, operating as part of the UNOSOM II force, attempted to capture Somali leader Mohamed Farrah Aidid in a raid. The ensuing battle resulted in 18 US troops, one Malaysian soldier, and between 300 and 1000 Somalis being killed; see Findlay (n 17) 200 for a detailed account of the incident.

[54] UN, *Comprehensive Report on Lessons Learned from United Nations Assistance Mission for Rwanda (UNAMIR) October 1993-April 1996* (New York: Department of Peacekeeping Operations 1996) 3.

[55] ibid 6.

safeguard international peace and security. Rather than allow their decisions to be affected by outside influences, the SC must ensure that future mandates, and the ROE that result from them, are based on events and situations on the ground where the operations are going to occur. Especially for the creation of ROE, the mandate operates only as an initial ceiling on how, and in what manner, force may be utilised as a tool by UN forces. IHL constitutes another layer of restrictions that must also be taken into account when formulating ROE.

B. The Influences that Shape the Creation of Rules of Engagement

ROE are developed according to the SC mandate and IHL to ensure that, when it becomes necessary, force is used in a manner that is controlled and legally justified. To understand the significance of the additional hurdle that SC mandates create for ROE development, it is also necessary to understand the minimum constraints on the legal use of force under IHL.

The former SG of the UN, Kofi Annan, published a bulletin dealing with specific principles of IHL that apply to UN forces. In it, he emphasises that the bulletin applies to UN forces 'when in situations of armed conflict they are actively engaged therein as combatants' and that they therefore apply 'in enforcement actions, or in peacekeeping operations when the use of force is permitted in self-defence'.[56] Although there was some confusion as to what 'situation of armed conflict' meant, the conclusion most in line with UN policy is that the principles specified in the bulletin apply at all times when recourse to force is utilised as an option by UN forces, regardless of whether or not the conflict in question rises to the level of an international armed conflict.[57] This represents a stricter application of IHL principles to UN forces, since the Geneva Conventions, with the exception of Common Article 3, apply only to situations of international armed conflict.[58]

With regard to the use of force, the bulletin sets out four general principles of IHL: (1) necessity, (2) distinction, (3) proportionality, and (4) humanity.[59] The first principle permits only that level of force necessary for achieving the specific mission goals.[60] According to the second principle, when UN personnel use force, they must always distinguish between civilians and combatants, and between

[56] *Secretary General's Bulletin* (n 19) s.1.1.

[57] Klappe, in Fleck (ed) (n 4) 625–626.

[58] *Geneva Convention Relative to the Protection of Civilian Persons in Time of War (Fourth Geneva Convention)* ('*Geneva Convention*') (adopted 12 August 1949, entered into force 21 October 1950) 75 UNTS 287 article 2.

[59] *Secretary General's Bulletin* (n 19) ss. 5–9; see, also, Cole et al (eds) (n 12) 5–6.

[60] *Secretary General's Bulletin* (n 19) 6.1; see, also, Jann Kleffner, 'Scope of Application of International Humanitarian Law' in Fleck (ed) (n 4) 59.

civilian objectives and military objectives, when targeting.[61] The third principle, proportionality, prohibits the use of force on a military target only when the incidental harm to civilians or civilian objects would be disproportionate.[62] Finally, the principle of humanity prohibits the infliction of any unnecessary suffering in the use of force; the use of force by the UN must be calculated and reasonable.[63]

While the majority of the principles enunciated in the Bulletin come directly from the Geneva Conventions and the Additional Protocols, the UN is not itself a party to any international treaties, even if the general principles of IHL are applied as a matter of policy.[64] In addition, the Bulletin states that the international treaties that participating nations are signatories to continue to bind the forces that they contribute, despite the UN having effective control over them.[65] Therefore, while technically the Bulletin will only bind UN forces in conflicts not amounting to an international armed conflict, if the situation does become qualified as such, the forces involved will be bound by any applicable international treaties and conventions to which their home-states are signatories.[66] In addition, any principles contained in those treaties and any conventions that have become settled state practice with the support of *opinio juris* would apply by way of customary international law, regardless of whether the nations in question are signatories thereto.[67]

The SC mandate is superimposed over these IHL principles. The influence of political and policy pressures on the mandate creation process, discussed in the previous Section, translates into ROE that do not conform to mission requirements. Clearly, the path from mandate creation to the official laminated Orders for Opening Fire card issued to each peacekeeper is not clear or direct and is influenced by much more than practical concerns for the efficient achievement of mission goals.[68] Force Commander Roméo Dallaire, for example, originally drafted interim ROE for UNAMIR that included permission to use force to defend 'persons under UN protection', but these were never approved, and the mission

[61] *Secretary General's Bulletin* (n 19) s 5.1; see, also, *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating the Protection of Victims of International Armed Conflicts (Protocol I)* ('*Additional Protocol I*') (adopted 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3 article 48; see, also, Kleffner, in Fleck (ed) (n 4) 60.

[62] *Secretary General's Bulletin* (n 19) s 5.4; see, also, *Additional Protocol I* ibid article 51(5)(b); see, also, Knut Ipsen, 'Combatants and Non-combatants' in Fleck (ed) (n 4) 94.

[63] *Secretary General's Bulletin* (n 19) s 6.3; see, also, *Additional Protocol I*, ibid article 35(2); see, also, Kleffner, in Fleck (ed) (n 4) 59–60.

[64] Michael Schmitt, 'Targeting in Operational Law' in Gill and Fleck (eds) (n 1) 246–247.

[65] *Secretary General's Bulletin* (n 19) s. 2.

[66] *Geneva Convention* (n 74) article 2; see, also, *Prosecutor v Dusko Tadic* (Appeal Judgment) ICTY-94-1-A (15 July 1999).

[67] *North Sea Continental Shelf Case (Federal Republic of Germany v Denmark; Federal Republic of Germany v Netherlands)*, [1969] ICJ Rep 3; see, also, *Statute of the International Court of Justice* (adopted 26 June 1945 as part of the *Charter of the United Nations*, entered into force 18 April 1946) article 38(1)(b).

[68] Holt and Berkman (n 40) 85.

was eventually mandated to only use force for self-defence.[69] Dallaire's draft ROE complied with IHL, but the mandate further restricted his ability to act.

5. The Great Divide: the Separation between UN Mandates and Rules of Engagement

Over-politicisation of mandates results in convoluted ROE that leave UN forces in morally and legally untenable situations. In some scenarios, the mandate freezes because UN forces become indecisive and hesitant to use force at all, regardless of whether or not they actually have the authority to do so. In other situations, there is an operational void between the role envisioned for the mission by the SC and the actual role that the mission is pushed into on the ground. The result is that UN forces are sent into conflict zones unequipped and untrained for the tasks they have to perform, and the parameters of their mission gradually 'creep' from a Chapter VI role towards a Chapter VII role, potentially violating IHL. After all, the legal foundations for the use of force are centred on the mandate, which makes any use of force outside of the parameters of that mandate illegal.[70]

These situations reflect a significant gap between theory and practice, as well as between decision-makers in the SC and those actually putting the mandate into practice on the ground.[71] UN mandates are unlikely to give direct guidance on what is actually expected for the implementation of that mandate, which makes the transition of authority for the use of force from the SC to the Force Commander unpredictable, especially considering how little input the Force Commander actually has in the creation of the ROE.[72] Parts A and B of this Section will discuss the two prevalent consequences that result from this dysfunction between mandate creation and ROE formulation: Frozen Mandate and Mandate Creep.

A. The 'Frozen Mandate'

Frozen mandates arise because of a number of factors but, in general terms, the causes can be grouped under two areas: (1) insufficient mandates and overly restrictive ROE and (2) inconsistent and uninformed interpretation of ROE. As was discussed in Section 3, in the cases of both the Irish contingent during the ONUC mission and the Belgian contingent during the UNAMIR mission, confusion over ROE caused peacekeepers to hesitate at the crucial moment in which they were killed. Even where the UN forces were militarily capable of defending themselves, overly restrictive ROE made the circumstances under which

[69] Astri Suhrke, 'Facing Genocide: the record of the Belgian battalion in Rwanda' (1998) 29 *Security Dialogue* 37, 44.

[70] Bruno Simma, 'NATO, the UN and the Use of Force: Legal Aspects' (1999) 10 *European Journal of International Law* 1, 6.

[71] Klappe, in Fleck (ed) (n 4) 629.

[72] Holt and Berkman (n 40) 84–85.

they believed they could use force ambiguous. Soldiers' unwillingness to breach the ROE, as they interpreted them, was arguably taken advantage of not only to kill the peacekeepers in question, but in the case of the UNAMIR mission, also between 500,000 and 800,000 innocent civilians.[73]

In SC Resolution 918 of 1994, the SC instructed UN forces in Rwanda to 'contribute to the security and protection' of civilians and recognised that its forces 'may be required to take action in self-defence'.[74] Based on this language, it is unclear whether the UN contingent would be legally permitted to use force if they were not fired upon personally; it implies that the mandate requires the UN force to decline to act.[75] Restricting peacekeepers' right to interpret hostile intent is an unreasonable limitation on ROE above and beyond the requirements at IHL. Instead, peacekeepers should be allowed to interpret both hostile acts and hostile intent. The failure to react robustly and consistently when it is justified degrades the credibility of the UN force, thus putting the entire mission at risk.[76]

When ROE are not permissive enough to allow UN forces to intervene in humanitarian crises when they are needed, the organisation created to stop atrocity simply becomes an observer. When Force Commander General Roméo Dallaire requested authorisation to seize weapons caches in Rwanda, after warning the UN Secretariat of the impending genocide, his request was denied because the mandate for UNAMIR did not permit such action; UN forces were only allowed to conduct weapons recovery operations to establish the weapon-free zone originally agreed upon in the Arusha Peace Accord. UNAMIR was forced to return the weapons they had seized to their owners. Conceivably, those same weapons were eventually used in the ensuing genocide, and Dallaire's force did not have the ROE or the manpower to conduct operations to stop it. This was what led General Dallaire to place his forces directly in the line of fire between civilians and Hutu militiamen. This would activate the self-defence principle if they were attacked, which allowed him to save thousands of civilians.[77] In the UN's report on UNAMIR, common themes are 'fundamental misunderstandings' of what UNAMIR needed for success, and 'false […] military assessments'.[78] In many cases, however, frozen mandates arise not out of incapacity but from misinterpretation of applicable ROE.

The failure of UN soldiers to use force even in self-defence shows that misinterpretation of ROE can be just as fatal as not having the required ROE at all. While the UN establishes its own ROE for each mission, soldiers taking part often have national ROE that may or may not comply.[79] The *Brahimi Report* attacked this phenomenon in particular, saying that the lack of common operating

[73] UN (n 70) 1; see, also, Findlay (n 17) 19.
[74] UNSC RES 918 (1994).
[75] Holt and Berkman (n 40) 85.
[76] UN (n 70) 3; Findlay (n 17) 371.
[77] Dallaire (n 36) 268–269.
[78] UN (n 70) 3.
[79] Cole et al (eds) (n 12) 5.

procedures, including interpretation of ROE, 'must stop' and that nations whose policies are contrary to those of the UN 'must not deploy'.[80] This statement by the SG represents a crucial call for standardisation of operation procedure for UN forces. Without it, UN missions lose legitimacy while civilians and peacekeepers alike are endangered. Such a situation arose in Rwanda, where certain UN contingents intervened to the best of their abilities to protect civilians, while others ignored what was happening around them. For example, General Dallaire believed that the Belgian contingent 'had serious misconceptions about the ROE, making them unnecessarily passive'.[81] This insight was, unfortunately, proven true when ten of the Belgian contingent lost their lives.

Similarly, frozen mandates can arise when national governments go around UN command structures and direct their forces while they are involved in UN missions without consulting force commanders.[82] In Srebrenica, the Dutchbat had the ROE to protect the Muslim refugees within their compound, but under orders from their national government, they turned them over to Bosnian soldiers, who in turn killed them. Recently, the Netherlands was held to be responsible by the District Court of The Hague for the deaths of 300 refugees that it turned over to Bosnian soldiers and whom they had the capacity to protect within their compound.[83] This follows the Court of Appeal of The Hague's decision in *Nuhanovic*, where it held the Netherlands accountable for three deaths arising out of the same incident and circumstances.[84] The Dutch state intervened, took effective control of the Dutchbat, and gave orders not to act that directly contradicted the orders issued by UNPROFOR headquarters. The need for unified and standardised ROE and command structure is readily apparent: implementation and interpretation, not inadequacy, caused the biggest problems for UNPROFOR's ROE.[85] Misinterpretation becomes an issue most often when ROE that are put forward as strict become fluid, such as when a mission's mandate gradually expands and changes from a Chapter VI to a Chapter VII operation.[86] While the change is made quickly on paper, formulating new ROE and having them applied uniformly by troops in the field while the operation is ongoing is much slower, and constant readjustment should therefore be avoided.[87]

[80] *Brahimi Report* (n 30) para 109.

[81] Findlay (n 17) 279.

[82] *Brahimi Report* (n 30) 45.

[83] Anna Holligan, 'Dutch State Liable Over 300 Srebrenica Deaths' *BBC News* (16 July 2014) <http://www.bbc.com/news/world-europe-28313285> accessed 17 August 2014.

[84] *Nuhanovic* (n 57); see, also, Klappe, in Fleck (ed) (n 4) 623.

[85] Findlay (n 17) 271.

[86] See, for example, the SC altering UNPROFOR's mandate through UNSC RES 807 (1993) to explicitly move the operation under Chapter VII for the first time while the mission was ongoing.

[87] Findlay (n 117) 373.

B. 'Mandate Creep'

Mandate Creep occurs when UN Force Commanders stretch the boundaries of the mandate for their mission in order to react to events occurring in conflict zones. While UN forces with insufficient ROE and badly formulated mandates make the UN look ineffective, UN forces that don't follow their mandate—or follow it haphazardly—run the more dangerous risk of making the UN seem incompetent.[88] When a UN mission begins to blur the lines between Chapter VI and Chapter VII by using force in ways that it is not mandated to do or is not authorised to do under its ROE, then it constitutes a breach of international law.[89] However, if a force is sent into a conflict zone with insufficient ROE to protect the civilians or safe-areas that they are supposed to protect, then the previously discussed scenario of mandate freeze arises. Unfortunately, the UN's solution to this problem has been to progressively expand a UN force's mandate piecemeal, reacting to instead of dictating events on the ground.

Mandate creep shows the same theoretical and political dependence on the principles of escalation of force and minimum use of force that UN peacekeepers are ordered to act under through their ROE. Reliance on these principles puts the mission at risk just as it puts UN forces at risk. In the former Yugoslavia, for example, UNPROFOR was sent into the conflict in 1992 in the absence of a firm cease-fire agreement: the force was lightly armed, could only use force in self-defence and was seen as an 'interim arrangement to create […] conditions of peace and security'.[90] This mandate was established in February 1992, and by September of that year, the SC passed Resolution 776, which allowed UNPROFOR to use force in order to protect humanitarian aid, but other than that the provision held UNPROFOR to the normal ROE of a Chapter VI mission.[91] This dysfunctional hybrid is representative of the hesitancy of the SC to recognise the conflict for what it was: an internal sectarian war with no peace to keep. It took until February 1993 for the SC to begin using Chapter VII language in its mandates and, as has been discussed in other incidents such as the Srebrenica Massacre in 1995, UNPROFOR was never able to consolidate and apply the ROE it was given so belatedly.[92]

The UN mission to Somalia shifted into a Chapter VII enforcement scenario as awkwardly as UNPROFOR had.[93] Similarly, UNOSOM I began operations in Somalia with no ceasefire to keep. The mission's primary purpose was to facilitate the delivery of humanitarian aid for victims of the civil war occurring

[88] ibid 370.
[89] Fleck, in Gill and Fleck (eds) (n 2) 111.
[90] Christine Gray, *International Law and the Use of Force* (Oxford: Oxford University Press 2004) 217–218; see, also, UNSC RES 743 (1992).
[91] Gray (n 175) 218; see, also, UNSC RES 776 (1992).
[92] Gray (n 175) 218–221.
[93] Findlay (n 18) 378.

in Somalia and began with a deployment of only fifty unarmed observers, which was later augmented.[94] UNOSOM I was not able to carry out its mandate and was supplemented by a completely new mission: UNITAF.[95] UNITAF was a multinational non-UN force that was mandated explicitly to act under Chapter VII to 'use all means to establish […] a secure environment for humanitarian relief'.[96] Thus, while acting under Chapter VII, the force was not to take offensive action against wrongdoers, but only to protect humanitarian relief. This, too, failed, and the mission was again redefined by the SC. Both UNOSOM I and UNITAF were replaced by UNOSOM II in 1993.[97] For soldiers taking part in the operations, the constant shift of mandate and ROE caused massive operational difficulties.

UNOSOM II had a broad mandate that called for it to prevent resumption of violence, secure disarmament of armed factions, maintain security at key locations—such as airports—and to assist in the protection of humanitarian aid.[98] While UNITAF was restricted to the use of force in self-defence, the UNOSOM II ROE went as far as allowing the use of deadly force without hostile action or intent in some situations. For example, 'crew-served weapons' were automatically considered a threat to UNOSOM II forces whether or not they showed hostile intent.[99] The same principle applied to 'armed individuals' within the areas under the control of UNOSOM II forces.[100] In addition, national ROE caveats further confused the situation, as each contingent used force in different degrees in different situations, with no standardisation to speak of.[101] UNOSOM II shows the devastating consequences of mandate creep from the peacekeepers' perspective, but also from the perspective of the host-nation population. The shift of missions from UNOSOM I, to UNITAF, to UNISOM II was confusing for its force members, but it is easy to imagine that Somalis likely saw no change at all; one day there were white foreigners and the next day they were still there. The only difference between the forces, or the ROE for that matter, for Somalis was likely their experience: one day they could carry weapons, while on the next armed soldiers pointed rifles at them and took their weapons away. The Battle of Mogadishu, in which eighteen Americans and an estimated 300 Somalis perished, marked the unfortunate climax of the confusion.

[94] UNSC RES 751 (1992).
[95] Gray (n 175) 222; UNSC RES 794 (1992).
[96] UNSC ibid.
[97] UNSC RES 814 (1993).
[98] ibid; see, also, Gray (n 175) 223.
[99] Findlay (n 17) 422–424.
[100] ibid.
[101] ibid 214–215.

6. The UN Mission in the Democratic Republic of the Congo: MONUC

The UN mission to the Democratic Republic of the Congo (DRC), MONUC, provides an interesting case study, as it represents a culmination of the topics discussed in this paper. While MONUC started off much as UNPROFOR and UNOSOM I did, with forces and mandates insufficient for its task, MONUC grew and developed in order to become one of the most vigorous and effective examples of the use of force by a UN mission in recent memory. It is unfortunate, however, that the mission had to start from behind and catch up as events developed. In 1999, the DRC was in turmoil, as civil war raged along tribal lines, as well as between foreign nations, such as Rwanda, that intervened in order to take advantage of the nation's rich natural resources. It is estimated that the conflict has claimed the lives of almost four million people.[102]

MONUC first took form as a small force of ninety military liaisons in 1999. Based on the findings of this advance mission, the African nations of the UN requested 15,000-20,000 troops with a robust Chapter VII mandate to help quell the civil war and unrest in the DRC and ensure implementation of the Lusaka Ceasefire Agreement.[103] Their request was denied. Still reluctant to send a force with such a proactive mandate, the SC fell back on its gradated approach of escalation of force and approved a force of 5,537.[104] Instead of the robust Chapter VII mandate that the African nations knew was necessary, the SC issued MONUC a mandate that resembled Chapter VI for the most part, with an element of Chapter VII grafted on at section 8 to 'take the necessary action […] as it deems it within its capabilities, to protect […] civilians from imminent threat of physical violence'.[105] MONUC began with a confusing and poorly drafted resolution, mandating the protection of civilians 'within its capabilities' but equipped it barely enough to defend itself, let alone apply its ROE proactively.

A. Same Mistakes, but Lessons Learned?

The size and uncertain purpose of the MONUC mission in 2000 meant that it was not able to operate effectively. In many cases, troops arriving for the mission were unaware of the dire conflict that they were entering and were untrained for the robust peace enforcement role that was expected of them. Different contingents interpreted the mandate and ROE in a variety of ways, and others were not even

[102] Coghlan et al, 'Mortality in the Democratic Republic of the Congo: a nationwide survey' (2006) 367 *The Lancet* 9504, 44–51.

[103] UNSC RES 1258 (1999); see, also, UN Doc S/1999/815 (1999), *Letter Dated 23 July 1999 from the Permanent Representative of Zambia to the United Nations Addressed to the President of the Security Council*; see, also, Holt and Berkman (n 40) 159.

[104] UNSC RES 1291 (2000) 3.

[105] ibid 4.

aware of the protection of civilians caveat in their ROE.[106] MONUC had the authority to use force to protect civilians as of 2000, but still took years to adjust because each contingent understood the mission in contradictory ways. As a result, the SC expanded MONUC in response to ongoing atrocity instead of preempting it.[107]

In 2002, the SC increased the MONUC mission to 8,700 troops and created two task forces to help with disarmament, demobilisation, and repatriation.[108] Even with these changes and the capacity to use force to protect civilians in their ROE, MONUC acted more as an observer mission. For example, in May 2002 MONUC had approximately 1,000 troops in the city of Kisangani, and yet failed to attempt to stop the massacres taking place there at the hands of RCD-Goma, a Congolese rebel group.[109] When fighting escalated in the Ituri province in 2003, the 700-strong MONUC force present there was incapable of protecting civilians in the area, due to the breadth of territory that the force was expected to patrol, leading the SG to comment on the 'immense gap between its capabilities and the high expectations of the population'.[110] As a result of the SC's early hesitation, the SG was forced to call upon France to lead an Interim Emergency Multinational Force (IEMF) in June 2003 to establish security where MONUC could not.

Operation Artemis, the operation led by IEMF, bought time for the SC to correct its mistakes, and set a firm example for when the mission was handed back to MONUC in September 2003. The French forces established a weapons-free zone and civilian protection area around the city of Bunia and enforced its control over that area aggressively.[111] The Ituri Crisis forced the UN to reconsider what the protection of civilians meant in the context of the deployment of a Chapter VII mission, and when IEMF handed control back to MONUC, the UN had organised the Ituri Brigade, which was comprised of approximately 4,800 troops, heavily armed and accompanied by combat helicopters. MONUC's troop ceiling was raised to 10,800, and its mandate expanded to 'take the necessary measures in the areas of deployment of its armed units […] within its capabilities' to ensure the security and freedom of movement of UN personnel, protect civilians from physical violence, and improve the security situation in the DRC.[112] To fulfil these goals, MONUC was authorised to 'use all necessary means to fulfil its mandate.'[113] This put the mission clearly within Chapter VII in terms of capacity, mandate and ROE. The difference in conceptualisation of the mission between pre- and

[106] Holt and Berkman (n 40) 191.
[107] ibid 159.
[108] UNSC RES 1445 (2002).
[109] Holt and Berkman (n 40) 159–160.
[110] Gray (n 175) 249; see, also, UN Press Release SC/7820 (2003), *Security Council Considers Way Forward in Democratic Republic of Congo, Hearing 28 Speakers*.
[111] Holt and Berkman (n 40) 162.
[112] UNSC RES 1493 (2003).
[113] ibid.

post-Operation Artemis MONUC is so great that it must be discussed in terms of MONUC Part 1 (MONUC1) and MONUC Part 2 (MONUC2). This separation is important, because MONUC2 represents the first concrete example of the UN successfully putting the *Brahimi Report* into practice.

B. The Use of Force and Application of ROE in the Congo

The use of force is, understandably, a last-resort scenario in any situation, but when it becomes necessary it must be reacted to quickly and assertively; MONUC2 is important because it shows that a UN force can walk the thin line between robust peacekeeping and war-fighting while maintaining the confidence of contributing nations, the host nation, and civilians. Clear mandates had noticeably positive effects on the outcome of MONUC2. Whereas once the SC spoke vaguely in terms of 'promoting a secure and stable environment', during the MONUC2 mission it clearly mandated the mission to 'protect civilians'.[114] The role does not come without costs, but when MONUC2 accepted its purpose and applied its ROE, it received resounding support from the SC, the UN in general, and, most importantly, civilians on the ground.

The formation and mobilisation of the Ituri Brigade was a strong and symbolic shift from the observation-reaction role of MONUC1 to the coercive-proactive stance of MONUC2. The deployment, however, was not flawless, and the Ituri Brigade still needed a sharp reminder of the more vigorous role that they were meant to take on. In early 2004, mutinous DRC troops occupied the city of Bukavu, and hundreds of civilians lost their lives. A typical scenario of mandate freeze occurred due to misinterpretation of shifting ROE. The UN deputy Force Commander, despite having access to attack-helicopters and troops on the ground, failed to utilise the assets at his command, or perhaps simply wasn't willing to put them in harm's way.[115] The result was a significant decrease in confidence and respect for MONUC2. The Uruguayan contingent that was in charge of defending the city became universally loathed by civilians in the area, leading some to comment that the 'Uruguayans [had not] come for peacekeeping, they [had come] for tourism.'[116] Thousands of civilians who had returned to their homes in the region after the Ituri Brigade had arrived once again fled.[117]

The events in Bukavu registered with the SC, however, as it approved an enlargement of MONUC2 to 16,700 troops.[118] Beginning in 2005, MONUC2 conducted some of the most aggressive peacekeeping ever seen. The force maintained uncompromising cordon-and-search operations to pre-empt attacks on

[114] Holt and Berkman (n 39) 187.

[115] Cammaert and Klappe, in Gill and Fleck (eds) (n 2) 155.

[116] James Traub, 'The Congo Case' *The New York Times* (3 July 2005) <http://www.nytimes.com/2005/07/03/magazine/the-congo-case.html> accessed 9 August 2014.

[117] Holt and Berkman (n 40) 164; see, also, Cammaert and Klappe, in Gill and Fleck (eds) (n 2) 155.

[118] UNSC RES 1565 (2004).

local villages, which by June 2005 had disarmed almost 15,000 militia members.[119] This, of course, attracted reprisals, as an ambush by a local militia, the Nationalist and Integrationist Front (FNI), killed nine Bangladeshi peacekeepers in February 2005. The reaction of Brigadier General Jan Isberg, the Commander of the Ituri Brigade, to these unfortunate events was important both for the success of the mission and restoring confidence in the assertive role that the UN was taking. Instead of backing off, General Isberg and the Ituri Brigade engaged the FNI during its cordon-and-search operations and killed fifty to sixty militia members during the ensuing firefights.[120] The 3,700-strong Pakistani contingent in particular was noted for its consistent application of the mission ROE to protect civilians. When it encountered Hutu rebel forces with links to the Rwandan genocide, the Pakistani contingent delivered the militia an ultimatum, and then moved in with aerial support to burn the camp to the ground; this was repeated with other armed groups on at least thirteen other occasions by October 2005.[121] It is likely that the Ituri Brigade's success was due to the fact that the Pakistani contingent that made up the bulk of its forces all followed the same ROE, both national and UN, and therefore did not have as many national caveats interfering with the mission. The international response to the proactive applications of MONUC2's mandate and ROE was positive, as MONUC2's mandate was further strengthened in March 2005. If coercive action was not expressly allowed before, it was spelled out in plain language in Security Council Resolution 1592, calling for increased cordon-and-search operations and coercive tactics.[122] This clearly shows the SC and the international community accepting and supporting the more robust application of ROE that MONUC2 had adopted. However, it did not occur without repercussions.

The expansion of the legal use of force in UN operations in situations that justify it is a step in the right direction for UN peacekeeping doctrine, but it must nonetheless be balanced against the repercussions that it causes. MONUC2 is, again, a perfect example. The assertive stance that the mission took had obvious positive effects in its areas of operation, as civilians gained confidence in the peacekeepers and a measure of security was restored. However, outside of the safe areas established by MONUC2, reprisal attacks against civilian populations took place. In addition, not all Non-Governmental Organisations ('NGOs') were willing to be affiliated with a UN mission that used force to achieve its objectives, and some pulled their support.[123] One of the primary purposes for the use of force by a UN mission is to provide for safe and efficient delivery of humanitarian aid to civilian populations. If the loss of support from humanitarian groups and backlash

[119] Holt and Berkman (n 40) 165.

[120] IRIN, 'DRC: UN Troops Killed 50 Militiamen in Self-Defence, Annan Says' (4 March 2005) <http://www.irinnews.org/report/53269/drc-un-troops-killed-50-%20militiamen-in-self-%20defence-annan-says> accessed 9 August 2014.

[121] Holt and Berkman (n 40) 166.

[122] UNSC RES 1592 (2005); see, also, Holt and Berkman (n 140) 165–166.

[123] Holt and Berkman (n 40) 157.

by militias outweighs the ground gained by applying mission ROE robustly, the decision to use force must be carefully analysed. These phenomena are foreseeable consequences of the proactive application of more permissive ROE, however, and if MONUC2 is an example for the future, then it has certainly vindicated the SC's policy decisions. In 2006, the DRC had its first democratic elections in 46 years, and in 2010, MONUC transitioned from a military role to a state-building role as it was renamed the United Nations Organization Mission in the Democratic Republic of the Congo (MONUSCO).[124]

7. Conclusion

In order for UN missions to be effective and compliant with IHL, the gap between the creation of mandates and the formulation of ROE for UN missions must be closed. The problems with the existing system have, unfortunately, come to the forefront through trial and error on an international scale. No single fix exists; however, MONUC has shown improvements that also serve as indicators of what solutions do in fact make a difference to mission efficacy when ROE are applied by soldiers on the ground.

UNPROFOR's experiences in the former Yugoslavia show that a Chapter VI mandated peacekeeping force cannot simply be incrementally adapted to become a Chapter VII mission. Not only does it risk infringing upon IHL by inviting mandate creep and inconsistent application of mission ROE, but it endangers the perceived impartiality of the force in general.[125] The relative failure and success of MONUC1 and MONUC2 reaffirm this lesson. MONUC1 was given a patchwork mandate that hinted at Chapter VII but remained Chapter VI, with a force that was hesitantly increased over time. MONUC2, on the other hand, was given a clear Chapter VII mandate with equipment and ROE to apply it. Another cause of misinterpretation comes from a lack of standardisation in ROE application, often due to national caveats that restrict soldiers' actions. The frozen mandate that results is equally dangerous for peacekeepers as for civilians and the mission. The SC must make efforts to standardise ROE before sending a mission to a conflict zone, and must refrain from scaling up the mission's ROE too frequently over the course of the mission.

The development of clear ROE for international missions comprised of multi-national contingents is a daunting task; however, it is essential in order for peacekeeping forces to remain effective in the twenty-first century. UN ROE will always be more restrictive than war-fighting, but must remain permissive enough to allow intervention where it is necessary; they must be flexible enough to allow local commanders to adapt to situations on the ground, but rigid enough to prevent mandate creep and the potential violation of international law. They must

[124] UNSC RES 1925 (2010).
[125] Gray (n 175) 227.

be sufficiently detailed so as not to leave room for error, but must also be clear and succinct enough to be memorised and carried by a soldier on a pocket-sized laminated card.[126] The bluff of international repercussions that worked so well for Lester B. Pearson during the Suez Crisis is no longer effective, and the UN has started to adapt to a new more proactive role maintaining international peace and security. While Rwanda, Somalia and the former Yugoslavia were devastating experiences, MONUC has shown that the SC and the international community as a whole have begun to accept the realities of modern peacekeeping.

[126] Findlay (n 17) 372–373.

The Divide over European Financial Regulation: an Economic and Legal Analysis of British Fears of Being Dominated by the Eurozone

Josef Weinzierl[1] and Lukas Koehler[2]

1. Introduction

THE EUROPEAN PROJECT is clearly at a crossroads. Not only do political and social developments challenge integration through 'ever closer Union' but further economic integration has been subject to very heavy criticism. As shown by the debate surrounding the UK referendum on membership in the European Union, one of the major points of contention lies in the field of financial regulation, in particular the protection of the single market.[3] One of the alleged underlying interest of the United Kingdom was to shield the City of London from increasingly burdensome regulation from Brussels, regarded as primarily serving the needs of Eurozone member states. What may be appropriate for the governance of the Eurozone was said to be a source of potential harm for the City. Reforms addressing this concern had been explicitly raised by David Cameron as an agenda item for negotiations to keep the United Kingdom in the European Union.[4]

This article seeks to determine the appropriateness of the United Kingdom's concerns and the calls for additional legal protection against the undermining

[1] Josef Weinzierl studied law at the University of Passau, Germany, and is a Magister Juris candidate at the University of Oxford.

[2] Lukas Koehler is a doctoral candidate at Bucerius Law School, Hamburg and a Magister Juris candidate at the University of Oxford. The authors would like to thank Anca Bunda for reviewing earlier drafts of this paper.

[3] Alex Barker, 'George Osborne makes shielding City priority in EU talks', Financial Times (London, 9 September 2015).

[4] David Cameron, 'Letter to Donald Tusk: A New Settlement for the United Kingdom in a Reformed European Union' (10 November 2015) <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/475679/Donald_Tusk_letter.pdf> (accessed 18 April 2016).

of the single market by Eurozone policy. Even post-Brexit, these issues remain relevant; the United Kingdom may well retain access to the single market. For this purpose, the diverging interests and concerns will first need to be defined and explained. Second, the mechanisms through which these interests may be realised will be analysed and put into the legal context. Thirdly, formal and substantive legal safeguards will be assessed against this background. Finally, we will conclude that only option to 'shield' the City is Brexit.

2. The Divide over Europe's Regulatory Integration

At the heart of the problem lies the separate existence of a currency union and a (much wider) single market. While the former is currently comprised of 19 members, the latter is currently composed of 28 members. The financial crisis of 2007/2008 and the sovereign debt crisis that began in 2010 showed that in an international financial world, risks cannot realistically be contained to single states, let alone financial institutions. The crisis was followed by rather modest reforms at the single market level. However, Eurozone states took on the ambitious project of building a banking union with the aim of significantly deepening regulatory integration.[5] Tension results from the fact that the Eurozone states—while pursuing deeper regulatory integration—are members of the single market at the same time.

3. The New Regulatory Landscape

A. Why European Financial Regulation?

Financial activity takes places across borders. There is no doubt that, even on an international level, banks cater for positive effects, such as the provision of a payment system, the diversification of risks for depositors, and the allocation of available short term capital to long term projects.[6] From an economic point of view, there is nothing in principle wrong with this as long as the economic risks inherent to the financial industry are not imposed on third parties. Banks' reliance on short term finance makes them dependent on the trust of market participants.[7] The sudden withdrawal of this short term finance may render their business insolvent. Creditors of banks B and C may take the difficulties of the first bank A as an indication that their banks will not be able to pay back loans, irrespective of the actual situation of B and C. Consequently, their creditors will also withdraw

[5] Cf. Communication from the Commission to the European Parliament and the Council, 'Roadmap towards a Banking Union', COM/2012/0510.

[6] Cf. John Armour et al, Principles of Financial Regulation (OUP, forthcoming 2016), Ch. 13.

[7] John Armour et al (n 6) Ch. 13.2.2.

their short term financing.[8] This leads to a vicious and indeed 'contagious' circle of reinforcing withdrawals of credit, ultimately endangering the viability of banks' business models and the real economy as a whole.[9]

The dangers for the economic system that come with financial activity materialise in the shape of either economic crises or state bailouts. Put differently, the financial risks taken on by individual actors may impose economic costs on society at large ('negative externalities' or 'costs'). To prevent these economic costs from arising in the first place, banks are subject to special regulation. Regulation seeks to internalise the costs into the banks' business models[10] by requiring, for example, a certain amount of equity to absorb losses.[11]

B. The European Dimension of Financial Regulation

Because of the interconnectedness of financial institutions in the single market, reference is often made to an internal market of financial services.[12] There is nothing special about the debt of a Finnish company ending up as an asset held by a Spanish bank. At the same time, the interconnectedness of the financial markets entails cross-border externalities.[13] In the same way that a financial institution may suffer from another bank's difficulties, financial systems in member states B and C may suffer from a financial crisis in member state A. This calls for regulation at the European level.[14] Two fora for regulation have to be distinguished: the single market on the one hand, encompassing all members of the European Union, and, on the other hand, the newly formed banking union.

1. The Eurozone: The Single Market

A common effort to regulate the financial industry is nothing new in the single market's legal framework. For instance, prudential regulation for banks has been harmonised through banking directives while, at the same time, leaving national legislators and regulators with wide discretion to transpose them into national law

[8] As an illustrative example for this mechanism in the onset of the financial crisis (within the shadow banking sector) see Gary Gorton, 'Slapped in the Face by the Invisible Hand: Banking and the Panic of 2007' (Federal Reserve Bank of Atlanta's Financial Markets Conference, May 2009).

[9] Cf. John Armour et al (n 6) Ch. 13.2.2.

[10] Cf. Anthony Ogus, Regulation: legal form and economic theory (Hart Publishing, Oxford, 2004) 35.

[11] John Armour et al (n 6) Ch. 13.2.3.

[12] Recital 7, Council Regulation (EU) 1093/2010 establishing a European Supervisory Authority (European Banking Authority), amending Decision No 716/2009/EC and repealing Commission Decision 2009/78/EC [2010] OJ 2 331/12 (henceforth EBA-REG).

[13] Guido Ferrarini and Luigi Chiarella, 'Common Banking Supervision in the Eurozone: Strengths and Weaknesses' (2013), ECGI Law Working Paper No 223/2013, 9 pp <http://ssrn.com/abstract=2309897> accessed 19 July 2016.

[14] Recital 1, 2 EBA-REG.

and regulatory practice. However, the Europe wide materialisation of systemic dangers in the shape of the financial crises laid bare the need to further deepen regulatory and supervisory integration at the European level. Consequently, reforms brought about significant substantive and institutional changes.

For the first time, the elementary provisions of banking regulation concerning own funds and liquidity were laid down in the harmonising and directly applicable Capital Requirements Regulation[15] ('CRR'), leaving no discretion to national legislators and regulatory agencies. This is complemented by the Capital Requirements Directive[16] ('CRD IV') concerned, *inter alia*, with: the governance of banks, a directive on deposit guarantee schemes[17] ('DGS') to provide for common rules on the insurance of bank deposits, and the bank recovery and resolution directive[18] ('BRRD'). The BRRD sets rules for the resolution of banks. These provisions form what is deemed the 'single rulebook of financial services'.

Equally significant are the changes to the institutional landscape. Member states agreed on the creation of the European Banking Authority[19] ('EBA'), charged with further harmonising the common rules and their application by provision of technical standards, guidance and the coordination of supervisory processes among national supervisors.[20] Unlike the European Central Bank, only in extreme circumstances is the EBA given direct supervisory powers vis-à-vis banks.

2. The Eurozone: A Banking Union

In the wake of the sovereign debt crisis of 2010, Eurozone leaders agreed on the creation of a banking union comprising three pillars. It comprises a single supervision mechanism ('SSM') that serves as a basis and precondition for a single resolution mechanism ('SRM'), and an envisaged common deposit insurance.[21]

This ambitious project responds to what can be described an 'implicit liability' for member states in the Eurozone: banks, at times, impose negative externalities onto the entire financial system. In order to avoid harmful economic distortions following from that, states often choose to bail out financial institutions on the

[15] Council Regulation (EU) 575/2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012 [2013] OJ L 176/1.

[16] Council Directive (EU) 2013/36 on access to the activity of credit institutions and the prudential supervision of credit institutions and investment firms, amending Directive 2002/87/EC and repealing Directives 2006/48/EC and 2006/49/EC [2013] OJ 2 176/338.

[17] Council Directive (EU) 2014/49 on deposit guarantee schemes [2014] OJ L 173/149.

[18] Council Directive (EU) 2014/59 establishing a framework for the recovery and resolution of credit institutions and investment firms and amending Council Directive 82/891/EEC, and Directives 2001/24/EC, 2002/47/EC, 2004/25/EC, 2005/56/EC, 2007/36/EC, 2011/35/EU, 2012/30/EU and 2013/36/EU, and Regulations (EU) No 1093/2010 and (EU) No 648/2012, of the European Parliament and of the Council [2014] OJ L 173/190.

[19] Cf. EBA-REG.

[20] Art. 8 EBA-REG.

[21] Euro Area, 'Summit Statement' (29 June 2012).

brink of collapse. Risks taken by individuals thereby end up in the hands of the public. Unlike members of a mere single market, however, member states of the Euro currency union do not have their own central bank and so only have a difficult route to financing by the European Central Bank ('ECB'). As a consequence, member states are pressured to leave the currency union. Such a step is feared to wreak havoc on the economies of other members of the Union, again inducing these states to indirectly assume responsibility for the first member states' debt. This is what happened with the European Stability Mechanism's ('ESM') financing of Greece, for example.

Common responsibilities, like access to ESM funds or the envisaged deposit insurance backstops, however, call for a common control that the EBA cannot provide.[22] Otherwise, banks have no disincentive to continue building financial risks, and national supervisors persist in overlooking these risks. This control is to be provided by the assumption of supervisory tasks by the ECB within the framework of the SSM. The ECB is given the task of directly supervising the Eurozone's 130 largest banks.[23] The ECB's remit also enables it to indirectly influence the national authorities' conduct towards the remaining banks by way of regulation, guidelines or general instructions.[24]

C. Diverging interests

Although the goals of regulation in the single market and the banking union are similar to a great extent, the level of integration is likely to diverge. This can be explained by the marked need for harmonisation within the banking union. Its success—the yardstick here being the economic survival of the currency union—depends on a stringent and coherent approach to financial regulation, which is likely to align all participants to act for the common goal. It also has the strong institutions necessary to achieve this goal.

1. 'Saving the Euro'

The sovereign debt crisis showed that there are implicit liabilities for all member states participating in a currency union. However, a common approach towards supervision alone will only help in respect of the specific externalities within the currency union. Chiefly, what may be needed is stringent regulation. It is quite clear that the ECB's rule making competences will only be confined to

[22] Philip Whyte, 'Britain, Europe and the City of London: Can the triangle be managed?' (Centre for European Reform Essays, July 2012) 6.

[23] Art. 6 section 4 Council Regulation (EU) 1024/2013 conferring specific tasks on the European Central Bank concerning policies relating to the prudential supervision of credit institutions [2013] OJ L 287/63 (henceforth SSM-REG).

[24] Art. 6 section 5 lit. a) SSM-REG.

administrative rule-making and might, therefore, not satisfy this demand.[25] This raises the issue of possible primary legislation, coming from the European level, that would be binding on all European single market participants, including the United Kingdom, to address this need. That way, the United Kingdom might be subjected to legislation tailored to Eurozone needs.

Banking union members as a whole have a strong interest in the coherent and effective regulation and supervision to remedy said negative externalities. Furthermore, such common control also serves as a moral and political precondition for access to common funds in shape of the ESM and the envisaged deposit insurance. Costs of badly drafted regulation and supervision of banks in member state A should not be borne by the uninvolved member states B, C and D. To attenuate the possibility for such 'moral hazard' behaviour of member states, a level playing field for financial services must be provided.[26] This is a project that may well translate into a legislative desire for very detailed and comprehensive harmonisation. The frequent mentioning of the 'single rulebook' being the 'backbone' of the banking union speaks volumes.[27]

2. Shielding the City

Although the situation is likely to change materially post-Brexit, the case of the United Kingdom still serves as an illuminating and representative example for the issue all non-Eurozone members of the single market face. In 2014, the contribution of the financial services sector to the United Kingdom's economy amounted to £127 billion in gross value added; around 8% of the total national output.[28] This is significant compared to Germany and France, whose financial industries only make up around 4%.[29] Beyond that, the export of financial services of the United Kingdom to the EU amounted to £20 billion, contributing about 1% of its GDP.[30] It follows that regulatory efforts impairing the financial industries' profitability for

[25] Art. 4 para 3 SSM-REG.

[26] Guido Ferrarini, 'Single Supervision and the Governance of Banking Markets' (2015), ECGI Law Working Paper No 294/2015, 3 <http://ssrn.com/abstract=2604074> accessed 18 April 2016.

[27] Cf. Council <http://www.consilium.europa.eu/en/policies/banking-union/single-rulebook> accessed 18 April 2016.

[28] Gloria Tyler, 'Financial Services: contribution to the UK economy' (House of Commons Library, 26 February 2015) <http://www.parliament.uk/briefing-papers/sn06193.pdf> accessed 18 April 2016.

[29] Statista, 'Anteil des Finanzsektors an der Bruttowertschöpfung Deutschlands von 1995 bis 2014' <http://de.statista.com/statistik/daten/studie/309545/umfrage/anteil-des-finanzsektors-am-deutschen-bip/> accessed 18 April 2016.

[30] Mark Hanrahan, 'Brexit: London Financial Sector Divided Over Risks of EU Departure' International Business Times (18 February 2016) <http://www.ibtimes.com/brexit-london-financial-sector-divided-over-risks-eu-departure-2311226> accessed 18 April 2016.

all member states will harm the United Kingdom's economy significantly more than others.[31]

This estimate is in juxtaposition to the supposed aim of financial regulation, namely benefiting the economy as a whole by the internalisation of negative externalities caused by financial activity. However, in a single market, those externalities are also imposed on parties sitting across borders, as the financial crises illustrate.[32] In that case, appropriate financial regulation will primarily benefit these countries' economies, while taking benefits from the City, previously enjoyed by the lack of internalisation.[33] The real impact of financial regulation can, ultimately, only be proved by reality. For the sake of the argument, this article henceforth assumes intrusive financial regulation may be detrimental to the United Kingdom's and other non-Eurozone member states economic interest.

In any case, this article is only concerned with what may be *perceived* as a threat in this regard. Calls for 'safeguarding' the City drew on the fear of other member states deliberately or recklessly harming the United Kingdom by introducing burdensome regulation at the European level, thereby damaging the country's economic model.[34] In general, this danger has always existed in the single market. The formation of the banking union, however, brings along a new level of common interest for strong regulation among participating members that makes this concern more severe.[35]

Having similar interests significantly facilitates coordination among actors. If the perceived costs that regulation brings are not evenly distributed among member states, unaffected countries have less incentive to abstain from harmful legislation. It was noted that Eurozone members have a common interest in strong regulation aimed at reducing moral hazard incentives and securing coherence of regulation and supervision.[36] In contrast, the United Kingdom's economic exposure to the financial sector attenuates its interest in strong financial regulation—at least as long as the United Kingdom's overall benefits from that regulation in the way that tax revenue from the financial sector minus the incurred costs from bail outs or economic crises is positive.[37] The bloc of 19 Euro countries may decide on regulatory projects among themselves (coordinated by the ECB or in Eurogroup

[31] Whyte (n 22) 4.

[32] Cf. John Armour et al (n 6) Ch. 4.6.

[33] This does not not necessarily imply that current regulation is softer than in other member states of the single market, cf. Whyte (n 22) 4, but given the integration of the banking union it might well be in future.

[34] Alex Barker (n 3).

[35] See Section 3.B.2.

[36] ibid.

[37] Cf. John Armour et al (n 6) Ch. 4.6.

meetings) and subsequently force this regulation through single market institutions. This phenomenon is what is referred to as the 'Eurozone caucus'.[38]

Not only could this facilitate the introduction of rules that harm the non-Eurozone members of the single market, it is also to be doubted whether their concerns will be seriously heard and taken into account at all, given that it will be outvoted anyway. As a consequence, the United Kingdom may lose its influence over the rules set for financial markets.[39] The United Kingdom has already lost the fight over rules on bankers' bonuses included in CRD IV and its further specification by the European Banking Authority. Remarkably, not a single non-Eurozone member state stood by the United Kingdom. A source of valuable experience and knowledge of policymakers and regulators, which could otherwise be harnessed to the benefit of the entire single market, may therefore in fact be left untapped.

4. The Potential for Discrimination in European Law-Making Procedures

The arguments above concern discrimination in the relevant procedures of financial law-making in the EU, which could, allegedly, result in detrimental and burdensome legislation for financial centers outside the Eurozone, such as the City of London and other non-Eurozone member states.[40] In an attempt to structure these concerns, it is illuminating to distinguish between two sorts of procedures. Firstly, there is the ordinary legislative procedure, which is especially used in Art. 114 Treaty on the Functioning of the European Union[41] ('TFEU')—allowing for a harmonisation of national rules aiming at the establishment and functioning of the internal market. Secondly, there are the voting rules in the agencies, especially the EBA (see Art. 44 EBA-REG), whose task it is to implement and complement the substantive banking regulation laws in daily practice. Both are potential sources of perceived discrimination from the perspective of non-Eurozone member states.

A. Voting in the Council of the European Union

The ordinary legislative procedure according to Art. 294 TFEU entails a qualified majority vote ('QMV') by the Council. The Single European Act in 1987 was the first instance where the member states—thereby amending the European Treaties—decided to depart from the unanimity requirement for voting in the

[38] Frank Vibert, 'Can Cameron achieve a new relationship between member states inside the Eurozone and those outside?' (BrexitVote, 20 November 2015), accessed 18 April 2016.

[39] Whyte (n 22), 6.

[40] See Section 3.C.2.

[41] This competence to harmonise has been widely used to establish the new supervisory authorities and substantive regulatory harmonisation in the past few years, see, for example, EBA-REG or CRR.

Council with regard to the Single Market and established a QMV procedure.[42] Yet, even under the Lisbon regime, there is an amended and rather complicated mechanism which tries to secure the interests of all member states while still allowing for efficient decision making, Art. 16(4)(5) Treaty on the European Union ('TEU').

The potential for discrimination arises from the fact that in the competent Ecofin Council,[43] the Eurozone member states have the necessary majority by themselves.[44] This makes it difficult for those countries which are not members of the Eurozone to influence the general framework of regulatory policy and the legal texts to a satisfying extent. Under the new mechanism established by the Treaty of Lisbon, which finally replaced the old mechanism on 1st November 2014, the 19 Eurozone members exceed the 65% qualified majority threshold and hence do not need the non-members to agree to their proposal. The assumption that this is, or could be, exploited remains speculative because it seems almost impossible that a proposal opposed by most of the non-Eurozone members would have a chance of success in the European Parliament. Furthermore, it is far from given that the Eurozone members would, as a whole, pursue a single regulatory strategy.[45]

However, the fact remains that the harmonised banking regulation rules, as incorporated in the CRR, are predominantly shaped and influenced by the Eurozone members and cannot be opposed by the non-Eurozone members, let alone by a single state like the United Kingdom. This of course reflects the current status of the EU in terms of integration because any member state can be outvoted in the Council. Whether there are effective safeguards against considerable harm to national policy by being outvoted shall be discussed in Part C of this Section.

B. Agency Rule Setting: The Arrangement in the EBA

Apart from the ordinary legislative procedure, there is a second formal source of discrimination identified by British fears. The main agency responsible for the implementation of the substantive regulatory laws, for soft law guidance, and the systemic stability of the financial system as a whole is the EBA, located in London. The EBA was equipped with a double majority voting rule, intended

[42] For an overview of the important changes brought about by the Single European Act see Maria G. Cowles, 'The Single European Act' in Erik Jones, Anand Menon and Stephen Weatherill (eds), The Oxford Handbook of the European Union (OUP 2012).

[43] Ecofin is the common abbreviation for the Economic and Financial Affairs Council. The Ecofin Council is the Council of the European Union in the special area of economy and finance. It is made up of the economics and finance ministers from all member states.

[44] See the detailed chart of the voting shares in British Bankers Association, Eurozone Caucasing, A Challenge to the European single financial market (June 2014) 14.

[45] Paul Craig and Menelaos Markakis, "The Euro Area, its Regulation and Impact on Non-Euro Member States, in Panos Koutrakos and Jukka Snell (eds), *Research Handbook on the Law of EU's Internal Market* (Edward Elgar, 2016) (forthcoming) 4(a)(i); also Vilbert (n 38).

to secure the interests of the member states without the Euro as their single currency.[46] When adopting a new standard or rule, the mechanism requires that there is sufficient support by Eurozone members and non-Eurozone members. It seems that British fears are less salient here because the non-Eurozone members succeeded in introducing this safeguard. Yet again, a single country cannot oppose certain initiatives of the authority if it does not find sufficient allies amongst its group. A pertinent example for this was the introduction of bankers' bonus caps promoted by the EBA, where the United Kingdom could not gather enough support for its opposition. Although the United Kingdom lodged an application of annulment at the European Court of Justice ('ECJ') as a result, it decided to withdraw the application once Advocate-General Jääskinen issued his opinion and recommended dismissal of the application.[47]

C. Discrimination as a Legal Test in a Voting Procedure

The specific form of discrimination that could arise in these circumstances is twofold. Firstly, the fear is that the ordinary legislative procedure allows the integrated Eurozone members to implement their view of regulatory policy by setting up rules which serve their interests. Secondly, the non-Eurozone members seem to assume that the EBA is a possible danger to financial centres outside the Eurozone as long as it orientates its regulatory policy along the interests of the Eurozone.

Legally speaking, both submissions raise the same formal question: is it possible that the outcome of the Union legislative process could technically amount to a discrimination of one specific (group of) member state(s) because the interests of the Eurozone members can be different from those of the non-Eurozone members? The question at the moment is not whether Art. 18 TFEU, which prohibits any discrimination on grounds of nationality, or any general principle of non discrimination under Art. 6(3) TEU,[48] covers this form of discrimination, but whether it is *ex ante* possible, once the respective procedure for the adoption of the legal text is followed.

[46] See Art. 44 EBA-REG. Note that it is almost impossible for the EBA to remain in London after the envisaged Brexit, which was already confirmed by EU officials in the aftermath of the UK Referendum.

[47] Case C-507/13 UK/Parliament and Council (ECJ, order of 09 December 2014); it is a rare case where the exception of Art. 51 Statute of the Court of Justice of the European Union applies so that the ECJ had direct jurisdiction in an action of annulment and not the General Court.

[48] This provision enables the ECJ, *inter alia*, to derive general principles of EU Law from constitutional traditions common to the member states. See, especially with regard to non-discrimination based on age, Case C-144/04, Mangold, [2005] ECR I-9981.

1. Discrimination by Being Outvoted

It is indeed hard to imagine that this question concerning discrimination can be answered in the affirmative because the European Treaties (TEU and TFEU) as an agreement of all member states ensure that the procedural and substantive interests of each of them are safeguarded to an appropriate extent in the respective fields of action. At the same time, there are material safeguards such as the principle of conferral or the principles of subsidiarity and proportionality, which are stipulated in Art. 5 TEU, that prohibit legal infringements, so that voting rules departing from unanimity as such cannot be exploited to harm disagreeing member states. In the area of interest at hand, one could point to the single market of financial services which is not to be impaired by measures that benefit the Eurozone, such as a deeper integration in banking supervision. Being outvoted in a majority decision cannot therefore amount to discrimination as such. A suitable mechanism depends on the member states as a whole. They set up a system where they find individual interests are sufficiently taken into account while ensuring a smooth decision making process.

5. A Perspective from the Principles of Proportionality and Subsidiarity

It is possible to reinterpret the concerns by focusing on the outcome of legislative procedures for the member states. The claim is that the EU imposes a regulatory concept that is dominated by Eurozone concerns and does not properly take into account the interests of a non-Eurozone financial centre like currently the City of London. Put this way, the claim is one of the vertical competence principles of subsidiarity (a certain measure has to be taken at the most appropriate level, for example the member state level) and proportionality (a European legislative or administrative act may not go beyond what is necessary to achieve the envisaged aim) as stipulated in Art. 5(4) TEU[49], preventing the EU from encroaching on member states competences in an unjustified way.[50] These principles intend to ensure that once the EU has a competence to act, this competence has to be exercised in a way that secures the competences which remain with the member states. An example is adopting a Directive rather than a Regulation because the

[49] The wording of Art. 5(3) and Art. 5(4) TEU is: 3. Under the principle of subsidiarity, in areas which do not fall within its exclusive competence, the Union shall act only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States, either at central level or at regional and local level, but can rather, by reason of the scale or effects of the proposed action, be better achieved at Union level.
4. Under the principle of proportionality, the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties.

[50] As most of the secondary law instruments are agreed upon using Art. 114 TFEU as internal market competence, both concepts apply to this shared competence, Art. 4 (2) (a) TFEU.

former generally, especially in its minimum harmonising form, gives the member states some leeway in implementing it. Applications challenging EU legislative acts based on such concerns became significantly more important with the introduction of a QMV system in the Council because those member states who did not support a proposal could challenge the proposal before the ECJ.[51]

Turned into a subsidiarity and proportionality claim, *viz.* that the planned regulative act is too intrusive and encroaches on the member states' area of competence, the interesting structural point remains that the argument does not rely on the regulatory autonomy of all member states potentially harmed by the harmonising legal act. Rather, it entails that individual member states are worse off because the impact on their financial sector is more severe than on other affected member states. However, the principles of subsidiarity and proportionality are designed to operate in order to maintain the vertical competence division as established by the TEU and TFEU and to prevent competence creep by the EU legislature, and not to balance the interests of individual member states in the political bargain.[52] It can be argued that the instruments of proportionality and subsidiarity review are not intended to legitimise political fears of those member states who did not convince the majority in the law making process. Instead, they are concerned with an objective, abstract assessment of the regulatory project in question with regard to an existing EU competence. As R. Liddle put it recently: 'A veto right for London on City questions would also breach a fundamental principle of the EU. If every member state demanded special protection for the sector which was most crucial to its economy, there would be no single market'.[53] This means that an interpretation of the principles governing European legislation as allowing individual member states to protect and veto whatever, in their view, is economically important would undermine the very concept of the integration, be it in the Eurozone or in the entire EU.

6. Safeguards against Potential Threats to National Interests

It is clear that the terminology of discrimination in relation to voting procedures is inappropriate. Nevertheless, it is illuminating to carve out the means by which individual or group interests of member states in the European legislature are protected and to assess whether these mechanism discussed below function as

[51] The most important case in this regard so far is Case C-376/98 Germany/Parliament (Tobacco Advertising I) [2000] ECR I-8419, because it is the only occasion the ECJ found Art. 114 not to be the correct legal basis and invalidated the legal act. In this case, Germany was outvoted in the Council and subsequently brought an action for annulment.

[52] See, Paul Craig, 'Subsidiarity: A Political and Legal Analysis' [2012] JCMS 72.

[53] Roger Liddle, 'Securing fair treatment between the "euro-ins" and "euro-outs"' (Policy Network, 6 November 2015) <http://www.policy-network.net/pno_detail.aspx?ID=4999&title=Securing+fair+treatment+between+the+%E2%80%98euro-ins%E2%80%99+and+%E2%80%98euro-outs%E2%80%99> accessed 15 January 2016.

effective means for safeguarding national interests, such as those of the United Kingdom in the financial regulation debate. In this respect, it is essential to distinguish between formal safeguards such as the discussed voting rules and substantive safeguards that constrain the action of the institutions.

A. Formal Safeguards: Voting Rules

Formal safeguards deal with procedural requirements in reaching decisions, as opposed to substantive mechanisms where the main tool at work is court scrutiny, and can be qualified as *ex ante* measures since they are used to ensure that the relevant interests are taken into advance. The most important formal safeguards when it comes to taking into account diverging interests at the European Union level are voting rules. Naturally, unanimity would be the strongest safeguard against any perceived discrimination because all interests would be accommodated. It is not surprising that the British government in 2011 proposed such a unanimity rule for voting that affected financial regulation rules.[54] The proposal failed as it did not attract enough support by other member states. The United Kingdom finally refrained from subscribing to the European Fiscal Compact.[55] The most important step with regard to voting rules certainly came about with the Single European Act 1987 which finally abandoned the unanimity requirement for agreeing on legislative proposals in the Council of the European Union for matters regarding the Single Market.[56] This marked a milestone in the integration process as it formed the transition from purely intergovernmental to a more supranational model where legislative acts could become binding on member states that did not support them, thus a major restriction on the sovereignty of each member state.

A weaker version of a voting rule safeguard has been implemented in the decision making process of the EBA, where according to the double majority requirement of Art. 44 EBA-REG Eurozone members as well as non-Eurozone members need to support the proposal to a significant extent, a simple majority of each group. Of course, voting rules in this procedure are not an efficient safeguard because the same lock does not exist in the law-making Council, so that the binding regulatory rules as such cannot be vetoed by the non-Eurozone members. Only the standards and technical rules which fall in the rather tightly constrained competence of the EBA can be vetoed this way.[57]

[54] ibid.

[55] For a detailed assessment see Michael Gordon, 'The United Kingdom and the Fiscal Compact: Past and Future' (2014) 10 European Constitutional Law Review 28.

[56] See Section 4.A.

[57] This is partly due to the Meroni doctrine of the ECJ; see Case 9/56 Meroni/High Authority [1958] ECR I-0011. For a thorough discussion of the current application of the doctrine, see: Merjin Chamon, 'EU agencies: does the Meroni doctrine make sense?' [2012] Maastricht Journal of European and Comparative Law 281.

There is a softer proposal to resolve issues of diverging interests in the decision-making process that would allow one member state to raise its concerns, the so-called 'Ioannina clause',[58] which essentially provides that each time a member state feels disregarded by its counterparts and raises concerns of national interest, it can elevate the respective topic from the Council to the head of states and governments in the European Council which then needs to find a solution. This idea was taken up in the European Council negotiations for the 'New Deal for the United Kingdom'. The decision—intended to take effect if the United Kingdom had voted to remain in the EU—contains a clause that enables a request for a discussion of the European Council on proposals which concern non-Eurozone members.[59]

In short, there are several formal safeguards in place to assuage the fears of non-Eurozone member states in relation to discrimination. Moreover, there would be at least four ways of implementing them in the EU legal order, each of them bearing different constitutional weight.[60] It is clear that the formal safeguards currently in place are not sufficient to reconcile the concerns raised by the United Kingdom in the debate of financial regulation but rather follow the integrationist path under way since the Single European Act. Thus, without significant changes, the procedures in the respective institutions in this sense do not provide an effective way of dealing with the fears of non-Eurozone members.

B. Substantive Safeguards in the European Treaties

The interests of a minority, such as the non-Eurozone members, do not have to be taken into account where there is a sufficient majority without those member states. The question that then arises is: what are the substantive legal safeguards for those minority interests, which protect them from being harmed in a legally relevant way and not just politically outvoted? The most obvious tool to scrutinise these legal acts is to challenge their legality by filing an action of annulment at the General Court under Art. 263(1), 256 TFEU, where member states, according to Art. 263(2) TFEU, are so-called 'privileged applicants'; they do not have to establish

[58] See Vibert (n 38). The origin of the term Ioannina-clause or compromise dates back to a Council meeting in this Greek city, see <http://eur-lex.europa.eu/summary/glossary/ioannina_compromise.html> accessed 18 April 2016.

[59] Conclusions of the European Council, EUCO 1/16, 19 February 2016. This new deal is discussed in great detail at by Paul Craig and Menelaos Markakis, (n 45). For the discussion of the Ioannina-clause (labelled 'emergency break') see ibid at 4(b)(v).

[60] Cf. Vibert (n 38), mentioning agreements (i) between members of the European Council, or (ii) between the institutions (Council, Commission and EP), or (iii) by a Protocol attached to the Treaties, or (iv) by changes to the internal provisions of the Treaties.

a specific legal interest in bringing proceedings.[61] The grounds for annulment are, however, limited to those mentioned in Art. 263(4) TFEU, so that general challenges seeking an overall assessment of a legal act are excluded. Furthermore, in judicial proceedings, the underlying policy concerns are not balanced again; rather, policy concerns are replaced by the court's assessment since this will not threaten the institutional balance. The long-standing jurisprudence of the European Court is to grant the EU legislator, as well as an expert decision-making body such as the EBA, a broad margin of discretion to reach specific policy decisions through the defined procedures.[62] Nonetheless, the Treaties place limits on this discretion which—regarding minority interests—primarily consist of the following principles, which act as constraints on the EU legislator.

1. Art. 18(1) TFEU: Non-Discrimination

Art. 18(1) TFEU provides that within 'the scope of application of the Treaties, and without prejudice to any special provisions contained therein, any discrimination on grounds of nationality shall be prohibited'. Art. 18 TFEU makes clear that the principle of non-discrimination forms one of the central pillars of the European Union's self-understanding. However, it is doubtful whether Art. 18 TFEU can be operationalised in the present context because the wording 'on grounds of nationality' as well as the systematic context of the provision, i.e. Part Two of the TFEU on Non Discrimination and Citizenship, indicate that the object of protection are the union citizens as individuals,[63] or the individual economic entities relevant for the free movement provisions. Art. 18 TFEU therefore seems to be a rather weak tool to challenge a legislative act with the argument that a member state was discriminated because his interests were not sufficiently taken into account. This terminology is questionable because of the very fact that the procedures in the TFEU preclude a direct discrimination from taking place. For example, the United Kingdom recently challenged a measure by the ECB claiming precisely that 'the ECB's location requirement infringes the principle of non

[61] In Case T-496/11 UK/ECB (CFI, 4 March 2015), where the UK successfully challenged the ECB's regulatory power to require clearing houses to be established in the Eurozone, this was discussed thoroughly as the UK is not part of the Eurozone. Despite contrasting reports, this decision contains no legal strengthening of the single market as opposed to an alleged policy of the ECB to discriminate the non-Eurozone member states, because the application was successful even before these matters were dealt with.

[62] See, for example, Case C-58/08 Vodafone and others [2010] ECR I-4999, para 52, pointing to the usually applied test that a measure has to be 'manifestly inappropriate'.

[63] Armin von Bogdandy, 'Art. 18 TFEU' in Eberhard Grabitz, Meinhard Hilf and Martin Nettesheim (eds), *Das Recht der Europäischen Union* (57th supplement 2015) paras 29 ff; Astrid Epiney, 'Art. 18 TFEU' in Christian Calliess and Matthias Ruffert (eds), *EUV/AEUV* (4th edition 2011) para 45.

discrimination in Article 18 TFEU'.[64] It is not clear from the judgment whether the argument is intended to point to discrimination of the member state or of the individual actors in the financial market in London since the Court did not have to assess this claim as it succeeded already based on other grounds. It is submitted here that only the latter claim is substantiated under Art. 18 TFEU because Art. 18 TFEU is explicitly addressed to the individual citizen of the European Union. Non-discrimination on the basis of Art. 18 TFEU can hardly be used with regard to an alleged discrimination of a single member state's interest in a specific policy *per se.*

In the context of financial regulation, one could imagine a potential discrimination of currencies other than the Euro, which is prohibited by the *lex specialis* to Art. 18 TFEU, the free movement of capital in Art. 64 TFEU, so that indirectly the Member State of this currency is harmed. This point is substantiated by Pavlos Eleftheriadis in the context of discussing the ECJ's judgment with regard to the ECB's location requirement mentioned in the last paragraph, which after the Brexit-vote presumably is legal history.[65] Such a perspective—although less relevant for the discussion of a possible discrimination of a member state *per se*—is illuminating and leads to the conclusion that the internal market is protected and hierarchically superior to Eurozone interests. Thus, potential discrimination of non-Eurozone members of the EU, which materialises itself in a less favourable treatment of the currency other than the Euro, falls under the pivotal free movement provisions. Therefore, legal protection against an act of the EU, or even of another member state, is available via the available mechanisms in the TFEU. Yet, in the context of common rule setting at the European level for example via the EBA, it is hard to imagine that a single currency outside the eurozone is discriminated in a legally relevant way.

2. Article 4(2), TEU: Equality of Member States

The 'Equality of Member States before the Treaties' is a primary legal principle, see Art. 4(2), TEU. It is intended to complement the protection of the national identities of the member states, stipulated in the same paragraph.[66] The ECJ stated in the context of new and old member states: 'The European Union is a union based on the rule of law, its institutions being subject to review of the conformity of their acts, *inter alia*, with the Treaty and the general principles of law. […] Those

[64] UK/ECB (n 61) para 78.

[65] Pavlos Eleftheriadis, 'The Proposed New Legal Settlement of the UK with the EU' (*U.K. Const. L. Blog*, 13th February 2016) <https://ukconstitutionallaw.org/> accessed 15 July 2016.

[66] On the interaction of national identity and equality in the context of the supremacy debate, see Federico Fabbrini, 'After the OMT Case: The Supremacy of EU Law as the Guarantee of the Equality of the Member States' [2015] German Law Journal 1003. It should be briefly noted that the threshold for 'national identity' is significantly higher than arguing about a regulatory policy in the City of London so that this provision is not discussed by itself.

principles are the very foundation of that union and compliance with them means, as is now provided for expressly in Article 4(2) EU, that the new member states are to be treated on the basis of equality with the old member states.'[67]

It is crucial to understand that equality in principle can be infringed in two ways: firstly, if like cases are not treated alike without objective justification, and; secondly, if unlike cases are treated alike without objective justification. It is only the latter possibility that is at stake in the debate about financial regulation rules that apply for the entire single market but can be agreed upon without the consent of the non-Eurozone countries. Although it is submitted that the equality mentioned in Art. 4(2) TEU is apt to safeguard smaller or less powerful states from EU law being exploited,[68] it does not act as a safeguard to being outvoted in a political agreement upon regulatory policies as long as EU law is not infringed. This underlines that *a priori*, the primary purpose of Art. 4(2) TEU in the context of equality is the first limb of the principle of equality mentioned above, i.e. that the EU shall be prohibited from treating member states differently where it is not foreseen by the Treaty.

Technically the equality requirement is closely linked with non-discrimination. One could describe these legal instruments as two sides of the same coin, non-discrimination being a negative prohibition and equality being a positive requirement. The useful peculiarity in our context of competing member states interests is that it is specifically the equality of these member states that is protected by Art. 4(2) TEU. As to the content, it is clear from the outset that equality before the Treaties already entails that the concept of equality does not require that all member states are always treated alike. Rather, a different treatment if set up in the Treaty is incontestable, for example the regulation for the allocation of MEPs for each member state in the European Parliament in Art. 14(2) TEU.[69] Another example can be found in the context of the single market for financial services. Art. 139 TFEU shows that the Treaty acknowledges the different status of Eurozone members and non-Eurozone members.[70] This contextualisation already indicates that it will be very hard to argue with the concept of equality of member states when trying to tackle legislative proposals that perhaps fit one of those categories more than the other, as long as the procedures which the Treaty requires are observed. Rather, as has been pointed out by Federico Fabbrini, equality argues for a uniform application of EU law because the equality is threatened by the very fact of allowing a single member state to derogate from it or raise national concerns in

[67] Case C-336/09 Poland/Commission (ECJ, 26 June 2012), paras 36 ff.

[68] See Armin von Bogdandy and Stephan Schill, 'Art. 4 TFEU' in Eberhard Grabitz, Meinhard Hilf and Martin Nettesheim (n 63) para 8.

[69] Example taken from von Bogdandy and Schill (n 68) para 7.

[70] See Walter Obwexer, 'Art. 4 TFEU' in Hans von der Groeben, Jürgen Schwarze and Armin Hatje (eds), *Europäisches Unionsrecht* (7th edition 2015) para 24.

varying contexts and thus allowing an EU à la carte.[71] Therefore, 'equality before the Treaties' cannot be invoked to demonstrate a breach of Union law just by being outvoted in the respective bodies.[72]

3. Article 114, TFEU and the Single Market

Article 114 TFEU was used as a legal basis for most of the new rules and agencies in the framework of European financial regulation.[73] One therefore initially has to assume that the substantial arguments of non-Eurozone countries such as the United Kingdom fall on fertile ground in Art. 114 TFEU. These countries substantially claim that the single market of financial services (freedom of movement for capital and services) may not be hampered by the deeper integration of the Eurozone in the context of the banking union and their majority in the Council as well as in the EBA.[74] A regulatory measure adopted under Art. 114 TFEU according to the ECJ 'must genuinely have as its object the improvement of the conditions for the establishment and functioning of the internal market'[75]. This conveys the impression that it operates as the best safeguard against discrimination of non-Eurozone member interests. There are several discrepancies in the approaches towards regulatory policies that should be pursued especially between the Eurozone and the non-Eurozone countries such as the United Kingdom.[76] The intention of Art. 114 TFEU clearly is to foster the internal market, and not a subset of it, such as the Eurozone.

Yet, for several reasons, it is difficult to construe Art. 114 TFEU as a safeguard against a specific regulatory policy that fits some member states better than others. First, the ECJ only once quashed a legislative proposal based on Art. 114 TFEU for lack of competence and in principle accepts the arguments of the EU legislature for which the ECJ developed a rather loose guide so that the threshold to fulfil the harmonisation criterion is lower than a first glance suggests.[77] Second, the focus is clearly on the single market and Eurozone concerns do not appear in a harmonisation measure based on Art. 114 TFEU. Nevertheless, the scrutiny of the Court—whether the correct legal basis was chosen and whether its requirements

[71] Federico Fabbrini (n 66) 1003, 1005; see, also, Case C-174/08 NCC Construction Danmark [2009] ECR I-10567, para 24.

[72] The same argument applies for Art. 4 (3) TEU. Under the assumption that the principle of sincere cooperation applies at all for the relationships of the (groups of) member states in a single EU institution like the Council or the EBA—which is more than doubtful—it is legally unimaginable to construct it as obliging the member states to take into account every national interest in the debate.

[73] See Section 3.B.1.

[74] See David Cameron (n 4).

[75] Germany/Parliament (Tobacco Advertising I) (n 51).

[76] The Bruges Group, The City of London Under Threat: The EU and its attack on Britain's most successful industry (The Bruges Group Publications, author: Tim Congdon) 15.

[77] Stephen Weatherill, 'The limits of legislative harmonisation ten years after Tobacco Advertising: how the Court's case law has become a "drafting guide"' [2011] German Law Journal 827.

are fulfilled—most likely does not reach out to an in depth scrutiny of the chosen underlying policy rationale. Regarding Art. 114 TFEU, it is essentially sufficient that there exist 'obstacles to the free movement'[78] which are then re-regulated on the European level and thereby replaced by European rules, such as in the Capital Requirements Regulation. However, because of the wide margin of discretion for the EU legislature,[79] it can hardly be expected that the Court finds a certain regulatory policy that was agreed upon under Art. 114 TFEU to be focusing too much on the needs of the Eurozone. Finally, it can be doubted that the single market can be accomplished if the concerns of one member state with a particular form of financial market such as the United Kingdom with the City of London are given greater weight in the shaping of legislative proposals than concerns of others. Thus, although the use of Art. 114 TFEU obliges the European legislature to set up harmonised rules apt for the entire single market, in practice this does not work as a safeguard against a specific policy pursued by the harmonisation measure.

7. Conclusion

This article has attempted to provide an economic and legal analysis of the implications of the financial regulation policy and the related legislative developments in the European Union, in particular with regards to the diverging interests between Eurozone 'caucus' and non-Eurozone member states. To illustrate this tension, we referred to the case of the United Kingdom. While the United Kingdom has an interest in promoting its internationally active financial sector, the Eurozone as a whole is, first and foremost, likely to foster strong financial regulation. This article also showed that concerns about possibly one-sided regulatory policy cannot amount to discrimination because of existing formal and substantive safeguards under European law. Non-Eurozone members' wishes to 'shield their financial industry' against the perceived threat of harmful European financial regulation are therefore hard to maintain under the current framework and can only be comprehensively realised by leaving the European Union altogether.

[78] See, for example, Germany/Parliament (Tobacco Advertising I) (n 51), paras 82–84 where the Court elaborated on the scope of the regulatory competence under Art. 114 TFEU.

[79] See Vodafone (n 62) para 52.

The Accession of Identical Chattels

Andrew Hill[1]

1. Introduction

WHAT IS THE proprietary result of two identical chattels acceding to one another? Imagine two javelin throwers competing at a range. Competitor One throws his javelin, which he owns, and it lands flat on the ground. Competitor Two then throws his javelin, which he owns, and, by some great misfortune, it impales Competitor One's javelin through the centre. Try as they might, they cannot pull the two apart.[2] Who now owns the resultant single chattel?

This article offers an answer. The word 'is' in the opening question is meant as a limitation. Established rules and concepts in property law will constrain my enquiry. Policy and fairness will have their role, but will not dictate the outcome. Those expecting a stunning normative proposition are advised not to hold their breath. Instead, this article attempts a conceptual response to a suggested rule for accession. The doctrine of accession will, indeed, be my second constraint: whatever rule results, it must be a rule *within* the doctrine of accession, not a replacement for it.

[1] BA in Law candidate, Lady Margaret Hall, Oxford. I would like to thank the editors of the *Cambridge Law Review*. The usual disclaimer, that any errors are mine, applies. On the occasion of this journal's inaugural edition, I would like to express that it is excellent to see this journal established to encourage and promote academic contributions from amongst the newer members of the legal community, and I hope that this journal has many interesting years to come.

[2] This inseparability might be more common if they were using the *pilum*, a heavy javelin used by the Roman army, the tip of which was designed to buckle on impact.

My purpose, in part, extends the concept of a case note to an analysis of two comments from secondary literature. They each last just two sentences and, while found in the same volume, are separated by almost 700 pages. They are the following, the first by Birks and the second by Hudson and Palmer.

> Suppose two sections of pipe or sheets of metal are welded together. In that case, if the pieces belonged to different owners, co-ownership of the resulting unit is the only solution.[3]

> If the two conjoined entities were of equal status so that neither could be regarded as principal or accessory, a situation not contemplated in the Roman texts, it has been suggested [citing Birks] that the solution should be ownership in common. The inactive party could, in principle, sue the improver in conversion and, on payment of damages, the totality would again become the improver's sole property.[4]

Taken in combination, these two passages propose that, where two chattels accede to each other, the outcome should be co-ownership of the resultant chattel. My purpose is to demonstrate that this proposition is problematic, and that instead another analysis is preferable. The plan for my argument is simple: Section 2 adds flesh to the presently bare bones of the proposition, Section 3 demonstrates the problems, and Section 4 contains my alternative suggestion.

[3] N Palmer & E McKendrick, *Interests in Goods* (2nd edn, LLP 1998) 238.

[4] ibid 932–933. This passage is not free from ambiguity. It was suggested to me that the passage proposes sole ownership for the 'active' party, being a party who plays a role in causing the accession, and that the 'inactive' party gets a claim in damages. I do not read Hudson and Palmer's passage to be saying this, however, for a number of reasons. First, discussion of awarding title to an active party, *qua* active party, sounds like manufacture, but the passage is written about the law of accession, which knows no such rule as awarding title based on involvement in procuring accession. Accession can result from a causal sequence which includes proximate human conduct, but does not necessarily require human input. Second, granting a personal award in damages because of a mechanism of property law is unheard of. It could still be a claim in tort, but abstraction means that, regardless of the operation of tort law, a pure proprietary outcome must also be devised in the common law of property. Thus, the law of property would not distinguish between tortfeasors and innocent parties. Third, for the 'inactive party' to have *locus standi* to bring an action for conversion which terminates his title, at which point the 'improver' becomes the sole owner, the inactive party must previously have had title, and that must have been as co-owner in order that the termination can elevate the improver's title to 'sole ownership'. Hudson and Palmer chose to cite Birks' proposition with no dissent, and thus seem to accept co-ownership as the correct outcome. Therefore, as I understand it, Hudson and Palmer envisage a scenario where A owns javelin X, B owns javelin Y. B deliberately does something to cause the javelins to accede. This results in co-ownership. However, because A does not want to be a co-owner, he brings an action in conversion either for the destruction of X as a separate chattel, or for some subsequent act of dealing by B. This action extinguishes A's co-ownership title.

2. The Proposition

It is useful to begin by contextualising the co-ownership proposition. Accession is the doctrine in property law which governs situations where two things physically attach to one another in such a way that one, the secondary chattel, loses its separate physical identity. If A's fence is painted with B's paint, then the paint accedes to the fence, A having title to the whole. If C's brick is used to build D's house, then the brick accedes to the house, D having title to the whole. In property law, this functions as a mechanism of destruction of property rights.[5] When the secondary accedes to the primary, its loss of identity equates to its physical destruction. Physical destruction entails the destruction of the title to it. The title to the primary chattel simply persists, the only change being the physical addition. Therefore, unlike the related doctrines of manufacture and mixture, accession is not a mechanism of acquiring rights. It is, and is only, a method of destroying rights.

Often, the operation of accession is easy to predict. If a difficulty were to arise, it would be for one of two reasons. The first, which is not presently concerning, concerns whether the degree of physical attachment between the chattels is sufficient to amount to accession. May it suffice to say for the present that the 'test' for accession of two chattels remains uncertain. English law premises the general test on two variables:[6] the degree of attachment, and the object (meaning 'purpose') of the attachment.[7] Both of the main authorities, however, concern chattels acceding to land, so the discussion is always carried out in parallel with, and hence in cognisance of, the factors relevant for considering fixture. Other tests have arisen in cases specifically addressing two chattels across the Commonwealth. One asked whether the chattels can be separated without destroying or seriously damaging either of them.[8] Another whether the things would be considered as having ceased to have a separate existence.[9] One case asked whether the separation would destroy the commercial utility of the chattels.[10] The most favoured test, however, holds that accession will occur through either loss of physical identity or through practical inseparability.[11] Whatever test is adopted, this is not our present

[5] B McFarlane, *The Structure of Property Law* (Hart 2008) 163–164; W Swadling, in A Burrows (ed) *English Private Law* (3rd edn, OUP 2013), 4.470–4.471.

[6] Note also the more specific indicia suggested by Lord Clyde in *Elitstone Ltd v Morris* [1997] 1 WLR 687 (HL).

[7] *Holland v Hodgson* (1872) LR 7 CP 328 (Ex Ch); *Elitstone Ltd v Morris* (n 6).

[8] *Bergoughan v British Motors* (1929) 20 SR (NSW) 61.

[9] *Per* Manning J (dissenting), *Lewis v Andrews & Rowly Pty Ltd* (1956) 56 SR (NSW) 439.

[10] *Regina Chevrolet Sales v Riddel* (1942) 3 DLR 159.

[11] *Rendell v Associated Finance* [1957] VR 604, 610; *Thomas v Robinson* [1977] 1 NZLR 385, 386–8; and now see *McKeown v Cavalier Yachts* (1988) 13 NSWLR 303.

concern. Herein, this article will assume in any discussion that there was sufficient attachment to amount to accession.

Our focus is instead on the second problem: identifying the primary and the secondary chattel. The authorities are anything but extensive. English law does offer a few basic rules, most notably that land is always the primary chattel.[12] Once again, however, the better authorities for accession of two chattels come from the southern hemisphere. In the New Zealand case of *Thomas v Robinson*,[13] the court held that items fitted to a car, including significant functional components like a new engine, carburettor and exhaust system, would accede to the car. The components are secondary; the body of the car is primary. In Australia, it was held in *McKeown v Cavalier Yachts*[14] that the components fitted to the hull of a yacht acceded to it, despite the components being worth significantly more than the hull. There was some suggestion that the result may have been the reversed, had they been fitted as one unit, rather than individually. Hence, the measure for determining the primary and secondary item is not value. Rather, the rule appears to be that whichever chattel contributes more significantly in determining the physical identity of the resultant chattel constitutes the primary chattel. Such thinking is evident in Roman Law, from which much of the applicable common law derives. Buildings and building materials accede to the land.[15] Corn accedes to land.[16] A purple thread always accedes to the garment into which it was woven, regardless of comparative value.[17] Writing accedes to the parchment.[18] There is, of course, the traditionally exceptional case of *picturae*, in which the canvas accedes to the painting, apparently justified by the 'policy' that a valuable painting should not be considered the secondary of the canvas on which it is painted.[19] Nevertheless, the *picturae* rule is consistent with the test of physical identity. An artistically painted surface obtains its form primarily from the image drawn upon it. When one, visiting the Louvre, sees da Vinci's *Mona Lisa* on the wall, one's primary reaction is to see it for the image it displays, not the object upon which it is drawn. Thus, the contribution to resultant physical identity rule appears to be implicit throughout the Roman law.

However, the situation contemplated in this article, the accession of two 'equal' or 'identical' chattels, causes that rule to falter. Two scenarios spell out the issue. First, two pieces of one metre copper piping lie end-to-end, one owned by Adam, the other by Bob. Somehow they become welded together, creating a

[12] *Hobson v Gorringe* [1897] 1 Ch 182 (CA).
[13] *Thomas v Robinson* (n 11).
[14] *McKeown* (n 11).
[15] D.47.1.7.10. (Gaius II *rer. cott.*); J.2.1.29.
[16] D.47.1.7.13. (Gaius II *rer. cott.*); J.2.1.32.
[17] J.2.1.26.
[18] G.2.77; D.47.1.9.1. (Gaius II *rer. cott.*); J.2.1.33.
[19] G.2.78; D.47.1.9.2. (Gaius II *rer. cott.*); J.2.1.34. One should note, however, that Paul gives a contrary opinion at D.6.1.23.3., in line with the rules for *scripturae*.

two metre pipe. Assume that the welding constitutes an accession. Who owns the two metre pipe? Second, as described in the introduction, two javelins become transfixed. Who owns the resultant object? The two contributions cannot be separated on the basis of which contributed more significantly to the resultant physical identity, because the contributions are equal. Nor could a test of value be applied. Therefore, it appears that the present law offers no answer to the question of which of two identical chattels is primary, and which secondary, and hence it cannot determine who holds the title, and who loses their title.

The two passages under consideration volunteer a solution to fill this legal void. Birks, writing in the context of fluid mixtures, asserts that co-ownership would be the *only* solution in the accession of identical chattels scenario. He acknowledges that the scenarios in question are not cases of mixture (or that, if they could be classified as mixture, the classification is tenuous).[20] However, predicting that the court would not be willing to apply the rules of accession either because they would not be willing to subordinate either chattel, he suggests that another rule is needed. Therefore, as a rule independent of both mixture and accession, but applying the outcome of the former[21] to a scenario involving the latter, he proposes the co-ownership solution. For this, he cites the Scottish case of *Wylie and Lochead v Mitchell*.[22] One might add, in passing, that the same solution would result in Switzerland[23] and Ethiopia,[24] but this appears to arise from the conflation of the principles of accession and mixture. Palmer and Hudson apply Birks' idea, seemingly, as a rule *of accession* rather than *in place of accession*.[25] They observe that Roman law offers no solution and, in the footnote, only repeat Birks' citation of *Wylie and Lochhead*. Therefore, their proposition is that Adam and Bob would, through applying the doctrine of accession, co-own the pipe, and Competitors One and Two would co-own the affixed javelins. I will now dispute the acceptability of this proposition.

[20] Palmer & McKendrick (n 3) 238. The mixture analysis is addressed below.

[21] Co-ownership is the result of a non-consensual mixture in English Law, whether by accident (*Buckley v Gross* 122 ER 213; (1863) 3 B & S 566; *Spence v Union Marine Insurance* (1868) LR 3 CP 427) or wrongful intention (*Indian Oil Corporation v Greenstone Shipping, The Ypatianna* [1987] 3 All ER 893 (Comm Ct)). If one accepts the Commonwealth authorities, a different rule may apply for consensual mixture: see *Farnsworth v Federal Commissioner for Taxation* (1949) 78 CLR 504; *Coleman v Harvey* [1989] 1 NZLR 723. However, since accession does not distinguish the outcome on the basis of consent/intention, this need not concern us presently.

[22] 1870 M 552.

[23] Article 727 of the Swiss Civil Code.

[24] Article 1183 of the Ethiopian Civil Code of 1960.

[25] Whether they really intend to call this a rule of accession, or just a rule related to accession, is immaterial for present purposes. I wish to question how *accession* would deal with the identical chattels scenario. The present article cannot, therefore, take Birks' liberty of avoiding the question, disapplying the doctrine and creating a different rule. It must find an answer within the law of accession regardless of how exactly Birks and Palmer and Hudson phrase their proposition.

3. Criticisms of the Proposition

There are three criticisms of the proposition. Each criticism assumes that it proposes a rule *within* the doctrine of accession, as this article seeks to find a rule within the parameter of that doctrine. They do not necessarily deny its validity absolutely. They do, however, call into question its conceptual coherence and normative desirability. Before developing these criticisms, however, two assumptions made by Birks must be remedied. First, he says that co-ownership is the only possible solution. Since this article intends to offer another, that assumption must be doubted. The second is that the doctrine of accession insists on ranking one chattel as primary, the other as secondary. While that is indeed its usual operation, in the absence of any specific authority on this point it is not an assumption one is entitled to make per se. With this in mind, let us proceed.

A. Strength of Authority

Wylie and Lochead v Mitchell[26] cannot serve as authority for the co-ownership solution within the law of accession. Instead, as Birks envisaged, it can at best stand for a separate proposition in place of the law of accession.

Messrs Wylie and Lochhead were funeral undertakers in Glasgow. They wanted a new hearse, and reached an agreement with Robert Hutton, a coachbuilder. He would build the main shell of the hearse, but they would supply the equipment and ornamentation which would be attached to it. In the end, their contributions cost £95 and £112, respectively. Mr Hutton undertook to build the carriage for them. He missed the completion date agreed and then, having nearly completed the work a few months later, went bankrupt. The work was completed at a minor additional expense. Wylie and Lochhead petitioned the trustees of Hutton's bankruptcy for the delivery of the hearse, claiming that they had the property in it. The trustees disputed their claim. The petitioners applied to the court, arguing before the First Division that the hearse was their absolute property, and that therefore they were entitled to recover possession rei vindications. They argued that either (i) the contract was not a sale, but rather a *locatio operarum* under which Hutton merely supplied his services and materials, or alternatively (ii) that it was a sale, but that property had passed by constructive delivery when the carriage acceded to the ornaments and equipment which they had supplied.

The court found that the contract was a contract of sale, so the matter turned on whether accession had occurred. A close reading of each judgment evidences a subtly different approach. Lord President Inglis held that accession does not apply, for two reasons. First, it was impliedly impossible to distinguish a primary and secondary chattel on the facts. Second, he considered that the (Roman) rules of accession are not always 'based on natural equity' or free of internal conflict,

[26] *Wylie and Lochhead* (n 22).

as Grotius identified regarding, for instance, the differing rules for *scripturae* and *picturae*.[27] He held that manufacture was also inapplicable.[28] Instead, his Lordship felt compelled to formulate a new principle according to natural equity, and natural equity only offered one solution: common ownership.[29] Lord Ardmillan considered that accession did produce an answer: the carriage was primary, the equipment and ornaments secondary. However, he considered that fairness demanded that this outcome, which admittedly was a narrowly drawn distinction on the rules of accession, was avoided. Instead, he held that the fairer solution was (impliedly) to treat this case as one of consensual mixture and manufacture, which he asserted led to the fair outcome of co-ownership.[30] Lord Kinloch impliedly acknowledged that the carriage was primary. However, he, like the others, felt that the all-or-nothing outcome of accession would be unjust because of the comparable values of the two contributions. This led him to assert common ownership as the fair and equitable outcome.[31]

Therefore, this case is no authority for a rule of co-ownership arising from the accession of identical chattels, for three reasons. First, accession seemed to offer a solution on the facts. Lords Ardmillan and Kinloch both considered that, were accession to be applied, the carriage would be the primary object. Therefore, the majority did not consider the chattels identical. Applying a test based on physical identity (not value, as Birks suggests), their Lordships considered that the carriage more greatly contributed to the end identity. Only the Lord President considered that it would be difficult to identify a primary chattel, and hence that the existing rules of accession were frustrated. Therefore, the majority contemplation of accession suggests that it could have applied in the ordinary way, with no additional rules being necessary. Second, the case may not factually be one of accession at all. It may be better analysed as one of combined mixture and manufacture, much like the Canadian case of *Jones v De Marchant* (albeit with a different outcome).[32] This analysis is adverted to, but rejected, in the Lord President's judgment. However, it appears to be the implied basis of Lord Ardmillan's judgment. Third, as already acknowledged and as Birks identified, the case is decided independently of, rather than upon, the rules of accession. All three of their Lordships make statements that they are not prepared to decide the case on the rules of accession because they considered that such an outcome would not be in accordance with natural equity. Therefore, they all locate the common ownership solution explicitly beyond the law of accession, rather than as part of it.

The first two reasons pose a problem for Birks. They demonstrate that this was not a case of identical chattels, or involving accession, at all. Birks locates

[27] ibid 557.
[28] ibid 556.
[29] ibid 558.
[30] ibid 561.
[31] ibid 564.
[32] (1916) 28 DLR 561.

his solution in a case which, on examination, never had to grapple with the problem. True, identical chattels may be an a fortiori case, because the issue in *Wylie and Lochhead* was with distinguishing incredibly similar chattels. That is not the problem, however. Instead, the rule from *Wylie and Lochhead*, when it is seen as applying in cases of distinct chattels, becomes more directly questionable. It operates as an alternative, seemingly available at the court's discretion when it considers the ordinary operation of accession unfair on the basis of the similarity of the chattels. Therefore, it risks undermining the general rules of accession. Instead of offering one rule, the law would offer two discretionary alternatives, a situation apt to introduce unpredictability. While one may respond that this need not concern us overly as the rule can only be invoked when the chattels are 'similar enough', Professor Birks would surely have been amongst the first to disavow such a vague and unpredictable threshold. A rule based on undefined or unqualified 'sufficiency' is no rule at all, but rather a vicious circle. 'When are two chattels sufficiently similar to invoke the *Wylie and Lochhead* rule?' 'Why, when they are sufficiently similar, of course!' *Wylie* and *Lochhead* neither concerns identical chattels nor offers a supportable rule, and hence is not good authority. As such, the proposition should not draw any strength from its reliance on this case.

We are left now in a legal wilderness. The one case mapping our path is no longer there to support us. The reasoning herein must, therefore, be theoretical. This commences with the two further arguments against the common-ownership proposition, one conceptual, the other normative.

B. The Conceptual Problem

The mechanisms of property rights and remedies, something with which Professor Birks engaged closely in his work, render the co-ownership outcome problematic. Rights arise from causative events. The causative events, according to Birks, are fourfold: consent, wrongs, unjust enrichment, and miscellaneous others.[33] Assume that the accession was not consensual. Granted, Adam and Bob could have consensually fused the pipes, but, seeing that there is a chance that it was not because accession is not inherently a consensual act (as one sees in the javelin case), assume that it was not consensual. Since accession is not inherently wrongful either—it can occur accidentally—assume that there is no wrong. There does not appear to be any miscellaneous causative event. While one might describe accession as a causative event in itself, this analysis is difficult to sustain, because, the question of the proper outcome for identical chattels aside, there is no acquisition of rights during accession (as explained above), and thus it does not function as a cause. The causative events are selected by a legal system as a matter of policy, and our

[33] Birks' literature on this is ample. See, e.g., *The Classification of Obligations* (Clarendon Press 1997); 'Rights, Wrongs, and Remedies' (2000) 20 OJLS 1; 'Personal Property: Proprietary Rights and Proprietary Remedies' (2000) 11 Kings College LJ 1.

system has followed the Roman rule of electing to treat accession as a mechanism of destruction alone. Nor can unjust enrichment operate to alter property rights, because the co-ownership solution must mean that the old titles are still destroyed by the accession in order for the titles as tenants in common to arise, and hence there is no transfer of title upon which to premise a claim in unjust enrichment. The factual gain in matter should not be sufficient without a transfer of title.[34] In any event, the common law only tends to award personal rights in response to unjust enrichment, so altering property rights is unlikely to be the proper response.[35] Thus, accession is not a causative event which can trigger rights, which is why accession is only a doctrine of destruction of property rights. Therefore, there is no explainable mechanism by which the co-ownership title can arise.

Nor should the co-ownership operation of accession draw support from the law of manufacture and mixture. Unlike accession, they both create and destroy rights. Non-consensual mixture results in co-ownership.[36] Manufacture grants title to the manufacturer.[37] Both instances also seem to lack the existence of a causative event, often involving very similar facts to accession. Therefore, one might reply, if these doctrines can create rights, why should accession not be able to? That argument rests on an assumption that the operation of mixture and manufacture is justified. This is not the place for a full assault on those doctrines. Instead, some basic observations will have to suffice. The title resulting from manufacture may be explainable without needing to identify a causative event in the process of manufacture. Manufacture results in a *nova species*[38] (hence why old titles to the materials are destroyed). As McFarlane observes, manufacture almost inevitably involves a person controlling the thing at the time or soon after the process is completed.[39] Therefore, it may be the case that the manufacturer owns the thing not because of the process of manufacture itself, but because of the simple operation of the ordinary *Armory v Delamirie*[40] rule for acquiring title by intentionally taking physical control of the chattel. Mixture is, however, problematic. There is no reason for a legal mechanism to operate at all, because the individual components in the mixture (if, perhaps, only at a molecular level) retain their original physical identities, and hence there is no physical alteration to the chattel necessitating the alteration of title. The only change is the creation of an evidential uncertainty.[41] The rule for mixture may need to be reconsidered, but that is for another time. It

[34] W Swadling, 'Ignorance and Unjust Enrichment: The Problem of Title' (2008) 28 OJLS 627.

[35] W Swadling, 'Rescission, Property, and the Common Law' (2005) 121 LQR 123, 135.

[36] See above (n 21).

[37] *Borden (UK) Ltd v Scottish Timber Products Ltd* [1981] Ch 25 (CA). The author will simply assert that the outcome in *Jones v De Marchant* (n 32) is erroneous and, in any case, not binding on the English Courts.

[38] A 'new thing'.

[39] McFarlane (n 5) 161.

[40] 93 ER 664; (1722) 5 Stra 505.

[41] See further the final paragraph of Section 3.B.1., below.

is sufficient to note for the present that, even if this evidential uncertainty could be treated as a miscellaneous causative event, it could not operate in the scenarios presently postulated, as it is evidentially clear who contributed which original chattel. Therefore, the rules of mixture and manufacture should not call into doubt the conclusion that accession involves no causative event.

Thus, the mechanisms of property law cannot explain how new rights, such as co-ownership, arise. The co-ownership solution is, therefore, conceptually problematic.

1. A Digression: Mixture Analysis

An entirely alternative analysis arising from this discussion of mixture could say that we should analyse the scenario in question *as* mixture. This raises a question into which English law fears to tread: what is the difference between accession and mixture? The commonest comment, prefacing any express endeavour to supply an answer, is that the borderline is difficult to define, and possibly fluid on a casuistic basis.[42] After that caveat, there normally follows one of two views. Hudson and Palmer seem to assume that the distinction is in terms of reversibility/separability: accession concerns physically irreversible unions, mixture concerns physically reversible but practically problematic unions.[43] Birks adopts a different view, in terms of the quality of the chattels: if the chattels are identical, it is mixture; if the chattels are non-identical, it is accession.[44] Since the subject matter of this article arises from Hudson and Palmer's application of the co-ownership outcome to identical chattels, it operates, aside from this sub-section, on the assumption that Hudson and Palmer have identified a supportable distinction. Nonetheless, the alternative analysis merits examination in passing.

Birks' view has much to commend it, but no decisive argument in its favour. It certainly fits a pattern. In mixture cases, it is normally two versions of the same product—oil,[45] jute,[46] tallow,[47] cotton,[48] and so forth—involved on the facts. Accession cases tend to involve different chattels. However, such facts equally accommodate Hudson and Palmer's distinction; the molecules in the fluid mixture cases, and the 'grains' (for the cases of jute and cotton, meaning 'bales') in the granular mixture cases are not physically bonded together. Conversely, the facts

[42] e.g. Palmer & McKendrick (n 3) 227–228.
[43] ibid 932–933.
[44] ibid 227–228.
[45] *Indian Oil Corporation* (n 21).
[46] *Sandeman v Tyzack* [1913] AC 680 (HL).
[47] *Buckley v Gross* (n 21).
[48] *Spence v Union Marine Insurance* (n 21).

involving non-identical chattels tend also to involve a physical bond. Thus, the cases sustain both patterns, so that cannot be decisive.

Birks may seem to have some assistance from authority, inasmuch as Staughton J in *Indian Oil Corporation*[49] seemed to premise the application of the doctrine of mixture on similarity of identity: 'where B wrongfully mixes the goods of A with goods of his own, which are *substantially of the same nature and quality*, and they cannot in practice be separated, the mixture is held in common'[50] (emphasis added). However, there are two issues with resting Birks' case on this lone sentence. First, the clause immediately after the added emphasis identifies separability—Hudson and Palmer's test—as part of the distinction as well, and hence it does not select one test. Second, Staughton J's judgment draws unmistakably on Roman principles,[51] yet it is not clear that a distinction in terms of identity of the chattels represents the Roman position. The Digest never claims to advert specifically to the issue of identical chattels.[52] However, at D.6.1.23.3., Paul comments that the welding of two identical materials results in loss of identity for the secondary material (an outcome consistent with accession), while in the case of soldering with lead (thereby introducing a different material), a different outcome would apply. This different outcome appears to be mixture, because he surmises that the *actio ad exhibendum* and then a *vindicatio* could be brought for the component which was attached by soldering. He says further that, in response to the *actio*, the soldered-on component could be detached, unlike the welding case. Thus, where the materials are identical, accession occurs upon welding, and the resultant chattel is inseparable. However, in the case of soldering, the materials are non-identical and mixture results, which is regarded as separable. Contrary to Staughton J's conclusion, therefore, this passage hints at a Roman distinction between accession and mixture based on separability, not identity. Thus, the initial arguments for Birks' distinction have less force than is prima facie apparent.

Hudson and Palmer's position also has some merit, primarily based on the practical oddity of the application of Birks' test. Identicalness would determine which doctrine applies in an occasionally surprising fashion. Physically bonding two circular copper pipes, one 100cm and the other 105cm, is accession, resulting in sole title for the owner of the 105cm pipe. Physically bonding two circular copper pipes, both 100cm, is mixture, and hence results in co-ownership. Small physical differences make for large legal differences. A still odder case is one of mixing sand. Suppose 100kg of black sand 'mixes' (in a lay, not legal, usage) with 100kg of white sand. The two are non-identical in terms of colour, chemical composition and, in all likelihood, sise and form, and so on Birks' test this is accession. However, one may have severe difficulty in specifying which is primary, and hence who is

[49] *Indian Oil Coporation* (n 21).
[50] ibid 360.
[51] P Stein, 'Roman Law in the Commercial Court' [1987] 46(3) CLJ 369, 371.
[52] Palmer & Kendrick (n 3) 932.

the owner. More helpfully, the sand scenario would be mixture on Hudson and Palmer's test as the granules are physically separable; separation is only practically problematic. Therefore, the separability test seems preferable. It also seems to have some Roman support. Gaius writes, '*sed et si sine voluntate dominorum casu confusae sint duorum materiae vel eiusdem generis vel diversae, idem iuris est.*'[53] The verb used for mixing is '*confundere*'. In Latin, this meant equally 'to mix', 'to pour' and 'to confuse'. The Roman terminology for mixture, therefore, had connotations of evidential uncertainty (a central mechanism by which practical inseparability can arise on Hudson and Palmer's test for mixture) implicit within it, a sense lost by the English rendering of 'mixture'.[54] Moreover, this passage accepts that the rules of mixture may apply to materials of a different nature—*materiae diversae*—which squarely rejects the Birksian identicalness test. Accepting this conclusion, applying the doctrine of mixture would not be the answer, as its application is not determined on a criterion pertaining to identity.

C. The Normative Problem

Returning to our main focus, co-ownership may also be normatively undesirable. A party may not want to find themselves a 50% co-owner of a chattel, because this could effectively lock them into the property. In *Wylie and Lochhead*, the Lord President noted that, 'such being their joint interest in a subject which is not capable of division, they must either bring it to sale and divide the proceeds in the above proportion, or the one must buy off the other by paying him the value of his proportion.'[55] Co-ownership is a common solution for mixtures because the mixture can be physically divided down into shares for each tenant in common to take. Co-ownership is common for land because land is capable of multiple simultaneous uses and is relatively permanent.[56] However, indivisible chattels are a different story. If Bob uses the two metre pipe, nothing much is left for Adam. But if Adam refuses to let him use it, Bob is equally likely to refuse any proposed use which Adam intended. They reach a stalemate. As the Lord President advises, they would have to seek a sale, either to their co-owner or to a third party. In principle, selling to a co-owner is a reasonable solution but, in practice, difficulties may arise. If one co-owner falls insolvent, the other co-owner will not be able to sell them

[53] (D.41.1.7.9. (Gaius II *rer. cott.*)). 'If, without the consent of the owners, two materials, whether of the same or different nature, have been '*confusae*', the rule is the same'.

[54] 'Mixture' derives from the alternative Latin verb '*miscere*', which possessed a more definite sense of 'mixture' or 'stirring', and is occasionally also used in the texts, e.g., D.41.1.7.8).

[55] *Wylie and Lochhead* (n 22) 559.

[56] Co-ownership of land does, of course, entail some issues. Clear examples of contention arise, for instance, where the parties refuse to cooperate over sale or possession. The resulting situation is now dealt with under sections 12–15 of the Trusts of Land and the Appointment of Trustees Act 1996, which have been the basis of much litigation. The distinction here is that, while land ordinarily has the potential to sustain multiple simultaneous uses, chattels normally do not.

their share. It is likely to be even more difficult to find a third party buyer. If the two present owners are at conflict over the use of a chattel, buying a share of that chattel seems to be an unattractive proposition. The problem of the undividable chattel is perhaps a reason why co-ownership has not been adopted as a solution in manufacture.

Hudson and Palmer suggest an alternative to sale in the second sentence of their quote, that one owner could sue the other for conversion under section 10 of the Torts (Interference with Goods) Act 1977, and hence, through a payment of damages, receive the value of their share and in turn lose their share of the title by section 5. However, this relies on one party being sufficiently active in relation to the chattel so as to commit a course of dealings which amounts to a conversion. If neither party is active enough (perhaps because they have both refrained from using the chattel until the dispute is resolved), then an action in conversion will not lie and, hence, this solution will be unavailable. Remember that, in this regard, many dealings which ordinarily amount to conversion will not be enough because, as co-owner, the defendant has a right to possession of the chattel. The only way he could commit conversion, therefore, is to deliberately deal with the chattel in such a way that excludes the other co-owner. Aside from destruction, transfer which successfully passes the full title and some more exclusive instances of use, no other instances appear to qualify. Thus, the action in conversion will not be widely available.

Furthermore, the damages for this conversion would be assessed at the price of the share in the new thing, not the old thing.[57] While this does not seem problematic prima facie, it may allow the active party to profit from his wrongdoing. Imagine the market value of one metre of copper pipe is £10, while two metres costs £16. When two pieces of copper pipe become joined together, if one applies Hudson and Palmer's suggestion, the inactive party who sues in conversion will only get his 50% share of the £16 back, not the full £10 (the value of the chattel

[57] *Kuwait Airways v Iraqi Airways (Nos 4 and 5)* [2002] 2 AC 883 (HL), [67]. However, it is recognised that, on occasions, the value of the award in conversion will be varied by the court: see, for example, *BBMB Finance v Eda Holdings* [1990] 1 WLR 409 and *IBL v Coussens* [1991] 2 All ER 133. If the court does permit such a variation, then this further objection is nullified.

he lost to the accession). Thus, the active party could profit from his conversion.[58] The inactive party can only sue for the present value of his share (£8). However, if the active party later physically separates the pipes into two, he will once again have £20's worth of piping. He makes an overall gain of £2, which, as explained above, cannot be remedied in unjust enrichment because there is never a transfer of title. He thereby profits from his wrongdoing. Therefore, because of its narrow availability and potentially inadequate remedy, the conversion solution is inappropriate. Thus, if one wishes to avoid lock-in and its potential consequences, one is advised to reject co-ownership as an outcome of accession.

4. Alternative Analysis

So, where does the answer lie? How can we formulate a rule for accession? The total efforts of Birks and Hudson and Palmer rest on the premise that the solution lies in triggering a new title. The solution here, however, looks to a different premise. It re-examines the factual analysis, from which flows an alternative legal consequence.

The attachment, I suggest, results in a *nova species*. Why? Since the original chattels are equal, the identity of each chattel changes by at least 100%. The one metre pipe undergoes a 100% increase into a two metre pipe. One javelin becomes two. The new chattel is as least as much different from the old chattel as it is similar. At most, the old chattel represents 50% of the identity of the new chattel, and this is likely to be less if the new chattel has additional characteristics not present in the old, such as a bend in the middle of the pipe through imperfect alignment where previously the two pipes were both straight. The more chattels that are involved, the more obvious this view becomes. If twenty planks of wood, all owned

[58] It was suggested to the author that such a wrong falls within Birks' scheme of potential causative events, and thus could justify an instance of acquisition, permitting co-ownership. Three observations should cast a sufficient shadow over this suggestion that it may be set aside. First, a response to wrongdoing which alters property rights can result in double compensation or double punishment (depending on how one would prefer to rationalise the action). There is likely to be a tort action—here, probably conversion—alongside. Thus, if property rights were altered in response to wrongdoing too, the wrongdoer would pay twice: once in a personal liability for damages, and once in the alteration of a property right. Hence, accession and mixture (see (n 21)) do not vary their outcome on the basis of wrongdoing. Second, the suggested rule for wrongful accessions could only be applied as an exception to the general rule of accession of identical chattels, as it can only apply in the context of wrongs, and many such accessions would not be wrongfully caused. Third, as Swadling notes (see (n 35) 136), there is no situation in which the common law responds to wrongs by granting property rights rather than an award of damages (subject to the power to revest in cases like *Car & Universal Finance v Caldwell* [1965] 1 QB 525, which Swadling there demonstrates is probably wrongly decided). The case for an interest in equity is also weak, as again equity normally responds in damages. Damages have been affirmed for third party liability in *Dubai Aluminium v Salaam* [2002] UKHL 48, [2003] 2 AC 366, and although there is increasing indication that a constructive trust may arise in cases of breach of fiduciary duty (e.g. *Attorney General of Hong Kong v Reid* [1994] 1 AC 324 and *FHR European Ventures v Cedar Capital Partners* [2014] UKSC 45, [2015] AC 250), there is a strong argument that these cases are wrongly decided (e.g. D Crilley, 'A Case of Proprietary Overkill' [1994] RLR 57; W Swadling, 'Constructive trusts and breach of fiduciary duty' (2012) 18 Trusts & Trustees 985).

by different people, somehow become joined together, it is hard to claim that any one of them plays the majority role in defining the physical identity of the resultant chattel. While the case seems weaker for two chattels because it is more plausible that one may play the majority definitional role, it is only a difference of degree. There is no apparent reason why one should treat the accession of two planks of wood or twenty planks of wood differently. Instead, the same conclusion should apply to both cases. Therefore, returning to a two chattel scenario, neither original chattel should be considered the majority contributor to the physical identity of the resultant chattel. Neither is primary. Instead, they may both be regarded as secondary things which lose their physical identity during the accession. Therefore, the resultant thing has no prior identity, and hence is a *nova species*.

This analysis depends, admittedly, on what qualities one prefers to emphasise when evaluating physical identity. However, some philosophical and linguistic guidance from our Roman counterparts may be usefully noted. *Nova species* is a concept most closely employed in the law of manufacture, stemming from the Roman doctrine of *specificatio*. Evident in the Sabinian-Proculian school debate over the proper proprietary outcome of manufacture was a difference in prior metaphysical philosophy. The Sabinians accepted the Stoic view of matter over form, hence why they held that the contributor of the materials should gain the title. The Proculians, however, followed in the Aristotelian and Peripatetic tradition which championed form over matter, hence awarding the creator of the new form—the manufacturer—title.[59] Justinian's basic rule for irreversible *specificatio* followed the latter tradition,[60] as has English law.[61] One should understand from this that one cannot simply assume that any one test is definitive of physical identity, as prior philosophical debate permits views to vary. However, the English legal tradition leans towards considerations of form over substance.

Van der Merwe has helpfully surveyed the Digest for the different tests applied in practice, and he isolates the three common verbs applied: *facere*, *transferre*, and *transfigurare*.[62] The concept of *facere* is unhelpful beyond manufacture, as the word is heavily premised in active human involvement, which accession does not necessarily demand. The best guidance is gained from the prefix '*trans*-'. Van der Merwe notes that these verbs have a sense stronger than that ordinarily seen for creating a *nova species*. Therefore, a high threshold for *nova species* was one which required that the things 'crossed' a boundary of physical identity. Our stronger case, that of twenty pipes or pieces of wood acceding, seems to fit this test; whatever one would describe twenty pipes roughly and chaotically latched to one another as,

[59] C van der Merwe, 'Nova Species' (2004) 2 Roman Legal Tradition 96, 100–101.

[60] J.2.1.25., adopted from what was apparently the opinion of Gaius, differing from his School (D.41.1.7.7 (Gaius II *rer. cott.*)).

[61] *Borden* (n 37).

[62] Van de Merwe (n 59). Although any direct attempt at translation will inevitably result in some degree of loss of the original sense, one can roughly equate these respectively to the English 'to make', 'to shift' or 'to transform' (in this context), and 'to transform' or 'to change form/appearance'.

'a pipe' or 'a plank of wood' is not the first term which comes to mind, not least because they will have probably lost their functional utility as simple pipes or planks. In any case, perhaps a more useful description of the ordinary threshold is gained, beyond a survey of verbs, from Ulpian's simple description *mutata forma*.[63] Whilst 'changed', 'altered', or 'modified' may be suitable translations, one cannot ignore the etymology of the English noun 'mutation' from *mutatio*. Could we say that two separate but equal chattels fusing to one another amounts to a modification or a mutation of their original forms? In ordinary English sense, one supposes that we could. Thus, one should consider oneself able to accept that a *nova species* does arise, even if doing so is initially metaphysically disquieting.

From these facts involving a *nova species*, the legal analysis arises. The ordinary rule that title vests in the first party into physical control applies. They become the absolute, highest title-holder. As for the other party, they may or may not be compensated. If the accession was consensual, they can make their own arrangements for remuneration. If the accession was wrongfully committed by the title-holder, they will likely be liable for the destroyed title in conversion, trespass and/or negligence. If the accession was accidental, there is a risk that they may not get compensation, but then the case for compensating them is weaker. Occasionally, property gets destroyed by pure accident, by natural causes and similar. Such are the risks of life.[64] Such also is the utility of insurance. If a person suffers a loss accidentally, so be it.

This analysis avoids the flaws of the co-ownership proposal. It does not rest on doubtful authority. Indeed, it rests on no authority at all. The case is conceptual. It avoids having to identify some absent trigger for new rights as a result of accession. It avoids co-ownership, so parties need not fear lock-in.

Is this the only solution? Not at all. I have not sought to demonstrate that the co-ownership solution is in any way inherently 'wrong', whatever that term would mean in this context. I have, however, sought to expose its flaws and offer an alternative analysis which may, in comparison, be preferable. One may, indeed, dislike both analyses, and, rejecting Hudson and Palmer's grounds of distinction between mixture and accession, step beyond the boundaries of accession and instead analyse these cases as instances of mixture (regardless of the semantic oddity of calling two welded-together pipes a mixture), and thus reach a co-ownership outcome by traversing a different path. None of the solutions are perfect, though I hope that any quibbles with my proposition will only be metaphysical dissents. And perhaps this prima facie imperfection was to some extent inevitable: all the theories have to override some accepted assumptions to reach their goal, because prima facie the common law is unprepared to tackle identical chattels.

[63] D.10.4.9.3. (Ulpian 24 *ed*).

[64] Those risks being reflected by the law in what Honoré terms '(risk-)distributive justice': see 'The Morality of Tort Law—Questions and Answers' in DG Owen (ed), *Philosophical Foundations of Tort Law* (Clarendon Press 1995) 73, 78–85.

5. Reflection

Having embarked on an almost untraveled adventure, we have taken an untrodden path, yet hopefully have reached our journey's end. Within the territory of accession, our paths remain two in number, though I suggest that the *nova species* analysis offers the less troublesome route. We may have no cases, and yet this does not prevent us from supplying an answer. And, when supplying this answer, I have suggested that we are cognisant of two things. First, one must ensure that one's solution is conceptually coherent. I hope to have demonstrated that the *nova species* analysis is coherent (indeed, it is not just coherent, but simple too), but that, for want of a causative event, the co-ownership analysis is not. Second, normative desirability must never be forgotten. Co-ownership risked lock-in, forcing parties into a potential proprietary stalemate through a process which may arise entirely naturally and accidentally. I do not think that the *nova species* analysis suffers from the same degree of normative deficiency, though I ask my reader to consider this for themselves. There may yet be alternative solutions, perhaps an entirely different outcome asserted on the basis of pure policy, or by side-lining accession and applying a different rule like mixture. Both of these alternatives likely depend heavily on the facts of any given case. In any event, it has not been my purpose to assess them, and I have not done so. Remaining within the bounds of established property law doctrine and confined within the law of accession, the *nova species* analysis should be sustainable.

When Equality Calls for Privilege: Sexual Assault and the Disclosure of Mental Health Records in Police Possession in Canada

Lauren Katz[1]

1. Introduction

FOR WOMEN WITH mental disabilities who allege sexual assault, privacy is inherently an issue of equality. This is particularly true for women who have had documented encounters with the police. Under the current record disclosure application process, the *Mills* regime, complainants' privacy interests are inadequately addressed, allowing mental health information to be sought by defendants on a discriminatory basis. The Supreme Court of Canada's confirmation in 2014 that police records are subject to *Mills*, and the 2014 release of a comprehensive police inquiry calling for increased police access to mental health information, jeopardise privacy and equality for sexual assault complainants with mental disabilities. A new class or statutory privilege between police and healthcare providers can protect complainants' equality and privacy rights while enabling a fair sexual assault trial.

2. The State of Sexual Assault and Privacy in Canada

It is estimated that in Canada only 0.3% of sexual assaults ever lead to a conviction.[2] In an assessment of attrition rates in sexual assault cases, only 3% of 460,000 sexual assaults from the past year were reported to the police and recorded as a crime. Of those assaults recorded as a crime, only 42% led to charges being

[1] Bachelor of Health Sciences, McMaster University, J.D. candidate (2017), Osgoode Hall Law School, York University. I would like to thank Osgoode Hall for encouraging students to publish their work and the *Cambridge Law Review* for extending this opportunity to students in Canada. I would also like to thank professors Michael Power and James Williams for fostering critical thinking in Privacy Law.

[2] Holly Johnson, 'Limits of a Criminal Justice Response: Trends in Police and Court Processing of Sexual Assault' in *Sexual Assault in Canadian Law: Leg Practice and Women's Action* (2012) 613, 632.

laid and only half those charges were prosecuted. Ultimately, only 25% of charges led to conviction.[3] However, the true percentages are impossible to know because it is believed that the vast majority of sexual assaults are never reported to the police.[4] Sexual assault is thought to be severely under-reported because victims often believe their privacy will be violated, that they will be scrutinised publicly, their personal health information will be used against them, and they will not be believed.[5] These beliefs are largely true. Defence counsel attempt to 'depict the sexual assault complainant as the irrational, incredible, and hysterical other of the rational legal subject,'[6] and aggressively pursue access to private documents through record disclosure applications to support an attack on credibility.[7] Privacy is violated through record disclosure processes in which 'boundaries of interiority are breached,' as when bodily integrity is violated in a sexual assault itself.[8] In a review of 48 record disclosure decisions in the first four years under the current third party record disclosure regime, the Department of Justice found that full or partial disclosure, or production of records to the defence, was ordered in 35% of cases; half of which involved records from multiple sources. The three most commonly produced records were counselling, medical, and psychiatric records, all of which attract a high expectation of privacy.[9]

While the violation of privacy in itself is problematic as a disincentive to reporting, it is also believed to be a major hindrance to rightful convictions. In the Department of Justice's case law review it was found that the grounds for seeking the production of complainants' records were based on prohibited myths and stereotypes about sexual assault in every single application.[10] Sexual assault and subsequent privacy concerns are clearly a gendered issue. Statistics Canada found that 92% of victims are female while 99% of perpetrators are male.[11] It is crucial that sexual assault also be recognised as an issue about ability. Women with disabilities are more likely to be sexually assaulted than other women, although the statistics are not conclusive.[12] Moreover, women with mental disabilities are particularly vulnerable to assaults on their credibility or capacity,[13] which

[3] ibid 630–632.

[4] Susan McDonald, Andrea Wobick and Janet Graham, Research and Statistics Division, Department of Justice Canada, *Bill C-46: Records Applications Post-Mills, A Caselaw Review* (2004) 14; Maire Sinha, 'Measuring violence against women: Statistical trends' (2013) 85 Juristat (Statistics Canada) <www.statcan.gc.ca/pub/85-002-x/2013001/article/11766-eng.pdf> accessed 10 April 2015.

[5] McDonald, Wobick and Graham (n 4) 14.

[6] Lise Gotell, 'The Ideal Victim, the Hysterical Complainant, and the Disclosure of Confidential Records: The Implications of the Charter for Sexual Assault Law' (2002) 40 Osgoode Hall LJ 251, 257.

[7] ibid 260.

[8] ibid.

[9] McDonald, Wobick and Graham (n 4) 24.

[10] ibid 40.

[11] Sinha (n 4) 29–30.

[12] McDonald, Wobick and Graham (n 4) 16.

[13] Janine Benedet and Isabel Grant, 'Hearing the Sexual Assault Complaints of Women with Mental Disabilities: Evidentiary and Procedural Issues' (2007) 52 McGill LJ 515, 542–546.

exacerbates their disadvantage once they report a sexual assault. The state of sexual assault and privacy in Canada strongly indicates that the privacy of sexual assault victims must be better understood. Once personal records are produced, the complainant's privacy is violated, regardless of whether those records contribute to the accused's defence at trial.[14] Therefore, the effects of third party record disclosure laws on complainants' privacy requires attention.

3. Overview of Record Disclosure in Sexual Assault Cases

A. Police and Crown Requirements to Disclose Relevant Information

The duties of the police and the Crown to disclose information in a criminal context are set out in two decisions of the Supreme Court of Canada: *R v McNeil* and *R v Stinchcombe*. The Supreme Court in *McNeil* ruled that the Crown has a duty to make reasonable inquiries into materials it is aware are in the possession of the police, which would be relevant to the defence or prosecution at trial. It also stipulated that the police have a duty to disclose to the Crown all relevant information pertaining to the investigation of the accused[15] and any other obviously relevant information.[16]

In *R v Stinchcombe* the Supreme Court determined that the Crown has a duty to disclose all relevant information in its possession, also known as the 'fruits of the investigation,' to the defence. This disclosure is subject to the Crown's discretion with respect to the relevance of the information, as well as the Crown's duty to protect privilege such as police-informer privilege.[17] Based on the disclosure made by the Crown, the defence may apply to have records produced. It is important to note, however, that the perpetrator is known to victims in 75% of reported sexual assaults,[18] so the accused is often already aware of records existing about the complainant and may proceed with an application on this basis.[19]

B. Access to Records Through Third Party Disclosure Applications

The accused can apply to the court to access records in the hands of third parties. The Supreme Court of Canada originally established requirements for third party record disclosure in the companion decisions of *R v O'Connor* and *A(LL) v B(A)*, which lay an important foundation for the later introduction of a statutory test. The Court set out the primary test for the disclosure of third party records in the

[14] Susan Chapman, Joanna Birenbaum and Janet MacEachen, *Factum of the Intervener: Barbara Schlifer Commemorative Clinic (R v Quesnelle)* (2014) [unpublished] [29]–[30].
[15] *R v McNeil* 2009 SCC 3, [2009] 1 SCR 66 [14] [*McNeil*].
[16] ibid [59].
[17] *R v Stinchcombe* [1991] 3 SCR 326, 1991 CanLII 45 (SCC) [*Stinchcombe*].
[18] Sinha (n 4) 30.
[19] Gotell (n 6) 274.

criminal sexual assault case of *O'Connor*. The test contemplated the relevance of the record and the need for the court to review records before producing them to the accused. The Court justified these components of the test on the basis that third party records are not in possession of the Crown and third parties have no obligation to assist the defence. Therefore, requiring the defence to prove the relevance of the record is a warranted shift in burden.[20] In *A(LL) v B(A)*, a civil sexual assault case, the Court determined that complainants and third party record holders can make submissions at disclosure applications and can appeal decisions to disclose.[21]

The *O'Connor* regime was replaced in 1997 when Parliament passed Bill C-46, which introduced ss. 278.1 to 278.91 of the *Criminal Code*. These sections present a comprehensive test for third party record disclosure in proceedings involving sexual offences ('the s. 278 scheme').[22] Parliament's intention in creating the s. 278 disclosure scheme was to engage in a contextualised analysis of the concerns with overcoming complainants' privacy rights in light of society's interest in reducing sexual violence against women and children.[23] Disclosure applications under s. 278 must follow a two-stage process.

First, the accused must prove that the third party record it seeks is 'likely relevant to an issue at trial or to the competence of a witness to testify' and that 'the production of the record is necessary in the interests of justice.'[24] At the first stage, s. 278.3(4) enumerates grounds that on their own are insufficient to support relevance. These grounds include the record's relation to the complainant's sexual activity and sexual reputation,[25] which are reminiscent of the prohibition of using sexual myths and stereotypes as a basis for defence in s. 276.[26] The list also forbids arguments for relevance based merely on the record's relevance to the credibility of the complainant, the reliability of the complainant's testimony because of the fact that the complainant has received psychiatric attention, and allegations of sexual abuse against persons other than the accused.[27] These latter three prohibited grounds are important when the complainant has mental health issues and has had documented encounters with the police.

If the defence can prove the likely relevance and necessity of production, the judge must then, at the second stage, review the relevant documents and decide whether to produce them to the accused. The judge 'shall consider the salutary and deleterious effects of the determination on the accused's right to make a full answer

[20] *R v O'Connor* [1995] 4 SCR 411, 1995 CanLII 51 (SCC) [31] [*O'Connor*].
[21] *A.(L.L.) v B.(A.)* [1995] 4 SCR 536, 1995 CanLII 52 (SCC) [27]–[28].
[22] McDonald, Wobick & Graham (n 4) 3.
[23] Martha Shaffer, 'The Impact of the Charter on the Law of Sexual Assault: Plus Ça Change, Plus C'est La Même Chose' (2012) 57:2d Sup Ct L Rev 337, 343.
[24] *Criminal Code*, RSC 1985, c C-46, s. 278.1.
[25] ibid s. 278.3(4).
[26] ibid s. 276.
[27] ibid s 278.3(4).

and defence and on the right to privacy and equality of the complainant,' and shall take into account the following factors:

> (*a*) the extent to which the record is necessary for the accused to make a full answer and defence;
> (*b*) the probative value of the record;
> (*c*) the nature and extent of the reasonable expectation of privacy with respect to the record;
> (*d*) whether production of the record is based on a discriminatory belief or bias;
> (*e*) the potential prejudice to the personal dignity and right to privacy of any person to whom the record relates;
> (*f*) society's interest in encouraging the reporting of sexual offences;
> (*g*) society's interest in encouraging the obtaining of treatment by complainants of sexual offences; and
> (*h*) the effect of the determination on the integrity of the trial process. [28]

Therefore, three *Charter* rights are invoked: the rights to privacy (s. 8) and equality (s. 15) for the complainant, and the right to make a full answer and defence (ss. 7 and 11(d)) for the accused.[29]

The constitutionality of this scheme was challenged in *R v Mills* on the basis that it violated the accused's *Charter* rights. The defence argued that the scheme violated the accused's right to make a full answer and defence, which is protected as a principle of fundamental justice under s. 7 in combination with the right to a fair trial in s. 11(d). The Supreme Court found the scheme to be constitutional because the scheme did not prescribe the extent to which an accused can access information in a trial. The scheme is prescriptive of a process rather than an outcome. As such, in order to be constitutional, the process must adequately account for all *Charter* rights affected.[30] The Court determined that the procedure outlined in s. 278 does indeed account for the rights to a fair trial, privacy, and equality comprehensively.[31] The fact that the scheme may have the effect of precluding disclosure to the accused, and therefore allow the Crown to access what the accused may not, is not itself an injustice, as long as the procedure by which this outcome is reached is a fair one.[32]

[28] ibid s 278.5(2).
[29] *The Constitution Act, 1982*, being Schedule B to the *Canada Act 1982* (UK), 1982, c 11 [*Charter*].
[30] *R v Mills* [1999] 3 SCR 668, 1999 CanLII 637 (SCC) [21]–[22] [*Mills*].
[31] ibid [139]–[144].
[32] ibid [116]. *Mills* was considered to exemplify a legislative-judiciary dialogue in the way it grappled with the differences between Bill C-46 and the *O'Connor* regime. See Frank Iacobucci, 'Reconciling Rights: The Supreme Court of Canada's Approach to Competing Charter Rights' (2012) 20:2d Sup Ct LR 137, 139–140.

4. Treatment of Police Records Under the Current Third Party Record Disclosure Regime

The Supreme Court of Canada in *R v Quesnelle* confirmed that police records are subject to the *Mills* regime. In *Quesnelle,* the Ontario Court of Appeal overturned the defendant's sexual assault conviction and ordered a new trial on the basis that a s. 278 application for disclosure of police records, which pertained to investigations unrelated to the crime being prosecuted, was dismissed in err.[33] The Crown appealed to the Supreme Court of Canada, where the conviction was restored. Justice Karakatsanis, writing for a unanimous court, notes that the *Mills* regime 'echo[es] this Court's frequent warnings against relying on myths and stereotypes about sexual assault complainants in assessing the relevance of evidence in the context of sexual assault trials.'[34] She then analyses the definition of 'record' in s. 278.1 of the *Criminal Code* to determine whether the police reports are 'records' subject to the *Mills* regime. The relevant components of the definition read as follows: 'record' means any form of record that contains personal information for which there is a reasonable expectation of privacy, but does not include records made by persons responsible for the investigation or prosecution of the offence.[35] The court held that complainants' police records were indeed subject to *Mills*.

The court undertook a two-part analysis to determine whether police records are captured by the s. 278 scheme, and if so, whether they fall under a statutory exemption. The first issue was whether there is a reasonable expectation of privacy in police records that would bring them within the definition of 'record' in s. 278.1. The jurisprudence on s. 8 of the *Charter* has clearly established that the expectation of privacy must be assessed based on the 'totality of the circumstances' and must not be restricted to only trust-like, confidential, or therapeutic relationships.[36] Police records create high expectations for privacy. They contain 'intimate personal information' that may 'do particularly serious violence to the dignity and self-worth of an affected person' if disclosed.[37] The risk of harm is two-fold: the complainant may be negatively affected by the disclosure of personal information to the accused for personal reasons, and the knowledge of such disclosure is a disincentive for victims to report sexual assaults.[38] The fact that a victim has disclosed assault-

[33] *R v Quesnelle* 2014 SCC 46, [2014] 2 SCR 390 [10] [*Quesnelle*]. First, the Court of Appeal found that complainants have no reasonable expectation of privacy in documents containing information they have given to police. Second, the court found that all police records prepared by the investigating police service are captured in the exemptions to the *Mills* regime specified in s. 278.1 of the *Criminal Code*. Applying the *Stinchcombe* regime instead, the Court of Appeal ordered the disclosure of the third party records and ordered a new trial.-

[34] ibid [17].

[35] *Criminal Code* (n 24) s 278.1.

[36] *Quesnelle* (n 33) [27].

[37] ibid [34].

[38] ibid [34]–[36].

related information to police does not negate their interest in privacy.[39] Therefore, complainants' police records do fall within the meaning of 'record' in s. 278.1. The second issue was whether complainants' police records are exempted from the records subject to *Mills* pursuant to the final words in the definition of 'record.' The use of a definite article in 'the offence' implies only records relating to the prosecuted offence can be exempted.[40] Additionally, the grammatical construction of the French language provision makes clear that the exception is for the records on the current offence themselves, and not the investigating officers making those records.[41] Furthermore, the purpose of s. 278.1 is to exclude only records that are so necessary to produce for a fair criminal trial that their relevance need not be discussed under an application, which does not logically encompass records of different incidents made by the same police service.[42]

In effect, the *Mills* regime must apply to police reports. This decision is certainly a success for complainants' privacy rights because the alternative—the Court of Appeal's approach—was to allow routine disclosure of unrelated occurrence reports to the defence. However, including police reports as records under s. 278.1 is still extremely problematic. Police reports raise a significant risk of discrimination to women with mental disabilities, and the *Mills* regime does not adequately address the right to equality.

5. Discrimination Against Women with Mental Disabilities in Police Record Disclosure

The disclosure of mental health records is fundamentally an issue of equality. Lise Gottell asserts that examining equality rights is essential to providing a full understanding of complainant's privacy interests, but '[t]o embrace a contextualised analysis linking privacy and equality, would deeply unsettle the individuated norms of criminal legal discourse.'[43] This is especially true when dealing with health information of complainants with mental disabilities. The next Parts describe how the issue of equality arises in the context of police interactions with women with mental disabilities and the disclosure of records relating to these interactions.

[39] ibid [37].
[40] ibid [46]–[49].
[41] ibid [50]–[53]. Since the English wording is ambiguous, the meaning conveyed by the French wording must prevail, as per *R v Daoust* 2004 SCC 6, [2004] 1 SCR 217.
[42] ibid [54]–[57].
[43] Lise Gotell, 'When Privacy is Not Enough: Sexual Assault Complainants, Sexual History Evidence and the Disclosure of Personal Records' (2006) 43 Alta Law Rev 743 [58].

A. Disadvantage at the Intersection of Gender and Disability

Gender and disability intersect to uniquely disadvantage sexual assault complainants.[44] Women with mental disabilities are not only more likely to suffer sexual assault than other women,[45] but also lead more heavily documented lives and are thus subject to greater privacy invasion in records disclosure processes.[46] Their lives are more heavily documented for various reasons that do not necessarily have any bearing on their credibility or competence to testify. For example, they may access multiple mental health services, rely on government, community, and police assistance, or experience one or more suicidal episodes.

Yet, mental health records are often sought under the guise of addressing the complainant's capacity to recount a sexual assault. The information is then used to undermine the complainant's credibility.[47] In other words, third party record disclosure applications are invoked on a discriminatory basis. Women with mental health issues are subject to third party applications based on stereotypes that they are untrustworthy and prone to lying.[48] Applications are 'unavoidably plagued by the stereotypes that women who report sexual assaults are 'crazy' or, where there is in fact a disability, that women with a mental or physical disability are unreliable.'[49] As discussed in Section 3.B, under s. 278.5(2) of the *Criminal Code* the courts are explicitly required to evaluate whether an application for record disclosure is grounded in a discriminatory basis or belief, and weigh this against other factors including the potential impact of the record on the integrity of the trial. As such, the statutory language implicitly recognises that there may be legitimate issues of credibility requiring record disclosure. However, the statutory language also suggests that the legitimacy must stem from information beyond the mere fact of mental disability and the assumption that it is categorically relevant. It

[44] Benedet and Grant (n 13) 519.

[45] McDonald, Wobick and Graham (n 4) 15.

[46] ibid 18–19; Benedet and Grant (n 13) 536.

[47] Benedet and Grant (n 13) 536–537. This is exemplified in the post-*Quesnelle* decision on admissibility of evidence, *R v A.G. and E.K.* 2015 ONSC 923 (CanLII), where the defendants were accused of intoxicating, raping, and abandoning a woman with a developmental disability. The trial judge found that two of the defence theories sought to be supported by the police record evidence essentially amounted to 'slagging of the complainant' and were unacceptable at trial. This is extremely problematic considering the judge on the s. 278 application had ordered the production of all of the complainant's police records relating to prior sexual assaults, in their entirety, to the defence, and yet only limited aspects of three of those records were actually deemed admissible. The admissible information related to prior sexual assault allegations that were 'demonstrably false' or 'recanted' as recorded in police occurrence reports. It is questionable whether the allegations were truly false or recanted since the complainant's developmental disability made her account of events imprecise and inconsistent. While her poor recount of events was heavily criticised at trial, it was not raised when the judge accepted that her 'recanting' of allegations was truthful. The defendants were later acquitted, primarily because of the complainant's unreliable testimony.

[48] ibid 539–540.

[49] Chapman, Birenbaum and MacEachen (n 13) [20].

seems that the court must address the possibility that the basis for record disclosure is a stereotype and no more. Yet, courts tend to ignore the topic of equality rather than actively engaging it in the discussion of privacy rights, as discussed in Section 5.C.

B. Perpetuation of Disadvantage Under the Mental Health Act

Changing policies and legislative provisions over the past 30 years have significantly expanded the police's role in the mental health system.[50] Most significantly, the *Mental Health Act* was introduced in Ontario in 1990 and prescribed a process for communication and collaboration between police and healthcare services. Under the *Mental Health Act*, the police may indicate that a person is in need of psychiatric attention, invoking a process by which that person is admitted to a psychiatric facility to be assessed.[51] This may happen to a woman who later becomes the victim of a sexual assault, and whose police records are sought for their 'relevance' to her credibility and competence, or to a woman who is 'emotionally disturbed'[52] at the time of reporting the sexual assault, in which case her mental state is recorded in the police occurrence report on the assault.

While this legislation is intended to enable a streamlined approach to people with mental illnesses, who pose a risk to themselves or others, it has the effect of empowering police officers to make 'lay diagnoses' of mental states.[53] Information from a mental health apprehension is then preserved in a police occurrence report that is typically accessible by any officer in the jurisdiction for years afterwards.[54] The police officer's decision that a woman is in need of psychiatric assessment, especially at the time of reporting a sexual assault, can undermine her credibility and form the basis of the defence's arguments later at trial. In effect, medically uninformed police impressions put women with mental health issues in the 'impossible position that the more marginalised and abused they are, the less likely they are to be believed in their initial and subsequent reports of sexual assault.'[55]

[50] Uppala Chandrasekera, *Police & Mental Health: A Critical Review of Joint Police/Mental Health Collaborations in Ontario* (Provincial Human Services and Justice Coordinating Committee, 2011) 2.

[51] *Mental Health Act*, RSO 1990, c M.7.

[52] This term is used by police to refer to people in crisis or people with mental disabilities, and includes those who are apprehended under the *Mental Health Act*, according to Frank Iacobucci, *Police Encounters with People in Crisis: An Independent Review Conducted by the Honourable Frank Iacobucci for Chief of Police William Blair, Toronto Police Service* (2014) 48, 73.

[53] Chapman, Birenbaum and MacEachen (n 13) [18]–[19].

[54] Chandrasekera (n 50) 42.

[55] Chapman, Birenbaum and MacEachen (n 14) [25].

C. Undermined Protection from Discrimination: the Mills *Decision*

The Supreme Court of Canada's reasons in *R v Mills* mitigate the impact of equality rights, even though the decision upheld the record disclosure scheme that purports to protect them. Lise Gottell criticises how *Mills* advances a 'highly individualistic and atomistic understanding of complainants' concerns.'[56] *Mills* reduces the list of factors the judge *shall* consider in s. 278.5(2) to a mere checklist of ideas the judge *may* consider.[57] This judicial sleight of hand removes the requirement to contextualise the complainants' concerns and consider the pervasive impact of violating complainants' privacy.[58] Furthermore, the list of grounds that are prohibited as the sole basis for a record application in s. 278.3(4), which are linked to discriminatory myths about sexual assault victims, is softened in *Mills.* The court claims that these grounds are actually permissible in some circumstances; the provision does not supplant the ultimate discretion of the trial judge reviewing the application.[59] Again, this weakens the very privacy and equality protections that the court defends as constitutional. *Mills* effectively presents a 'contest between privacy and fair trial rights, conceived in a zero-sum manner' that 'becomes the only focus of judicial analysis' while equality is relegated to the background.[60]

Indeed there is ample evidence that privacy is construed narrowly by judges as a direct and individualistic antagonist to the right to make a full answer and defence, thus diminishing the significance of equality rights.[61] For example, a research report published by the Department of Justice found that, out of 39 post-*Mills* decisions on record disclosure applications, 29 cases mentioned the accused's defence rights and 28 cases mentioned the complainant's privacy rights. Only four cases engaged in an analysis of equality rights. Furthermore, the influence of discriminatory beliefs or biases, a factor listed in s. 278.5(2) that directly speaks to equality, was only mentioned in 20% of cases in the review.[62]

D. Further Privacy Issues: Calls for Increased Police Access to Mental Health Information

A likely increase in police involvement with the mental health system has the potential to exacerbate the differential documentation and disbelief of women with mental disabilities. An independent inquiry into the Toronto Police Service's (TPS) encounters with people in crisis, conducted by the Honourable Frank Iacobucci, overwhelmingly advocates for even greater involvement of police in the mental health system. In light of recent trends towards deinstitutionalization and freedom

[56] Gotell (n 43) [27].
[57] ibid [29]; *Mills* (n 30) [134].
[58] Gotell (n 43) [29]–[30].
[59] *Mills* (n 30) [120].
[60] Gotell (n 43) [43].
[61] McDonald, Wobick and Graham (n 4) 31; Benedet and Grant (n 13) 539–540.
[62] McDonald, Wobick and Graham (n 4) 31.

to decline medical attention for people with mental disabilities, many of these people end up encountering police in crisis.[63] Several police services throughout Ontario have developed guidelines and programmes to obtain mental health information, with consent, to better serve people with mental health issues.[64]

The Toronto Police Service report suggests police officers should have greater access to mental health information. The report recommends the development of a protocol 'to allow the TPS access to an individual's mental health information in circumstances that would provide for a more effective response to a person in crisis.'[65] The report continues to list relevant privacy factors that should be addressed in the protocol and contemplates the possibility that police be included in a patient's 'circle of care.'[66] The circle of care consists of health service providers for a particular patient who can share information about that patient freely in order to provide coordinated care effectively.[67] Therefore, if police are included, they could have access to patient information without the patient's consent.

The report anticipates a need for written agreements between police and psychiatric facilities regarding patient rights, including privacy rights.[68] However, even effective privacy protections will nevertheless fail to address the issue of equality. Criticisms of the current record disclosure process suggest that the more health information police can access, the greater the risk will be for complainants with mental disabilities that their police records are sought and used to discredit them at trial.

6. Promoting Privacy and Equality with Police-Healthcare Privilege

Given the failure of s. 278 to do justice to equality, the balancing act in s. 278 applications appears to, at best, precariously protect the privacy of police-documented women with mental disabilities. For these women to have an equal right to privacy while allowing the police to expand their role in the mental health system, their privacy must be protected from the outset by default.

Privilege makes privacy the default. Establishing a privilege over communications between police and psychiatric facilities for the purpose of enhancing the mental health system would better protect these complainants' privacy than relying on the unpredictable application of the *Mills* regime to police records. Pursuant to *Stinchcombe*, privileged communications made known to the Crown are not to be disclosed to the accused, unless their privilege is challenged

[63] Iacobucci (n 52) 83.
[64] Chandrasekera (n 50) 39–41.
[65] Iacobucci (n 52) 111.
[66] ibid.
[67] Chandrasekera (n 50) 42.
[68] Iacobucci (n 52) 104.

and deemed an unfair limit on the right to make a full answer and defence.[69] The effect that privilege has on preventing disclosure is that a s. 278 application would be less likely to be initiated. If it were initiated, the privileged content would at least attract an extremely high expectation of privacy, which would factor into the judge's analysis of the application. Although it is not entirely impregnable, privilege sets the highest threshold possible for overcoming privacy rights.

Privilege is a logical response to the issues of equality and privacy for female complainants with mental disabilities for two reasons. First, discrimination based on myths about gender and disability, leading to violation of privacy rights, is already manifest at the record disclosure stage. Equality and privacy would be more effectively and predictably protected at an earlier stage: from the initial creation and sharing of mental health information by police. Second, equality requires that where complainants are uniquely disadvantaged by their disability and police encounters, these confounding factors must be controlled by putting these complainants on equal footing with all other complainants. Essentially, the defence should have to argue for records' relevance without capitalising on the fact that the complainant has had a mental health-related encounter with police. Requiring the defence to challenge privilege in order to gain access to police records shifts the discussion away from the mere facts of disability and police encounters and towards the realm of real relevance.

There are three broad categories of privilege, two of which may suit police-healthcare correspondence: class privilege and statutory privilege. Class or *prima facie* privilege is the recognition of privilege for an entire category of communications at common law, which includes solicitor-client privilege and informer privilege. Statutory privileges are legislated, such as the statutory religious privilege in the *Quebec Charter of Human Rights and Freedoms*.[70] The third category, case-by-case privilege, recognises privilege on an ad hoc basis at common law.[71] This form of privilege is unpredictable, much like trends in protecting privacy in post-*Mills* record disclosure applications.[72] It is unlikely to be helpful in reducing unnecessary disclosure to the Crown as it can only be established at trial, which has failed to recognise privilege even between therapists and patients or between priests and penitents in some instances.[73]

A police-healthcare privilege would be unique from other recognised privileges in that it would protect communications and exchanges of information between the two named parties for the benefit of third parties—people with mental disabilities. However, it would also provide police and healthcare practitioners the benefit of open and honest communication with the knowledge that they will not harm the

[69] *Stinchcombe* (n 17).
[70] *R v Gruenke* [1991] 3 SCR 263, 1991 CanLII 40 (SCC) [*Gruenke*]; *A.(L.L.) v B.(A.)* (n 21) [37]–[39].
[71] *A.(L.L.) v B.(A.)* (n 21) [39].
[72] McDonald, Wobick & Graham (n 4) 31.
[73] *Gruenke* (n 70).

people they seek to help by sharing their health information. Police officers may be more reluctant to take note of and access mental health information if they find their records are being aggressively pursued to discredit sexual assault complainants and defend the perpetrators their colleagues have arrested. Likewise, healthcare practitioners may be reluctant to share more mental health information with the police if they find that the information negatively impacts their patients when they are the victim of crime. Privilege would protect from these harms to enable the police and healthcare practitioners to fulfil their responsibilities to people with mental disabilities confidently. This concept is somewhat analogous to the solicitor-client privilege in that the communication is protected because the lawyer needs full disclosure from the client in order to serve the client's legal interests as best as possible. Here, the police and healthcare professionals need substantial disclosure from each other to best address mental health needs.

It is important to note that the s. 278 scheme already presumes records pertaining to sexual offences cannot be disclosed to the defence.[74] Privilege does not change that. Privilege instead has the effect of limiting the police's disclosure and production of any mental health information in its possession to the Crown, from whom its existence would have to be disclosed to the defence pursuant to *Stinchcombe*. Even if privileged communications become the subject of a s. 278 application, the privacy and equality interests will be much more powerful in relation to the accused's right to a full answer and defence.

A. Establishing Class Privilege

The possibility of establishing a new class privilege for private records relating to sexual assault complainants was contemplated and ultimately rejected in *A.(L.L.) v B.(A.)*. The minority judgment, delivered by Justice L'Heureux-Dubé, provides a comprehensive framework for deciding when a class privilege is appropriate. It also sets a precedent for finding that private records of complainants—in that case, arising from the therapist-patient relationship—could meet at least some of the criteria for establishing such a privilege. While weighing the benefits and disadvantages of recognising a class privilege, Justice L'Heurex-Dubé highlighted four principles governing when a class privilege can be established at common law: (1) the privileged relationship must be inextricably linked to the justice system; (2) the privilege must be justified by compelling policy rationales similar to those that support the solicitor-client privilege; (3) the privilege must be ascribed to a narrowly defined class; and (4) granting privilege must not infringe the truth-seeking process

[74] *Criminal Code* (n 24) s 278.2(1).

at trial.[75] While the prospective therapist-patient privilege failed to satisfy the third and fourth principles, police-healthcare privilege could satisfy all four principles.

1. Inextricable Link Between Police-Healthcare Relationship and the Justice System

In *A.(L.L.) v B.(A.)*, Justice L'Heureux-Dubé drew upon the Supreme Court's reasons in *R v Gruenke,* which provided that new class privileges should be inextricably linked to the justice system.[76] She found that there was an inextricable link between the therapist-patient relationship and the integrity of the criminal justice system because complainants' awareness that their personal health information could be disclosed would logically deter them from seeking treatment and contribute to under-reporting of assaults.[77] The Supreme Court had already recognised that 'chronic under-reporting of sexual assault cases undermines the effectiveness of the criminal justice system.'[78] Similarly, the fear of having mental health information disclosed to the defence can also deter women with disabilities from engaging with police in the first place, and deter health professionals and police from engaging frankly with each other, as discussed at the beginning of Section 5.

Furthermore, the police-healthcare exchange of mental health information affects the administration of criminal justice at the prosecution stage. *Quesnelle*'s affirmation that the s. 278 scheme applies to mental health records in police possession supports the conclusion that the police-healthcare relationship is connected to trials and pretrial disclosure. As discussed in Section 4, the prosecution of sexual assaults is greatly undermined by disproportionate access to complainants' mental health information in a system where reporting and prosecution of sexual assaults are already woefully low. Given the recommendations of the Toronto Police Service inquiry report, police are likely to encounter even more mental health information, incidentally putting more personal health information at greater risk of being exposed to the defence. The discussion of therapy records in *A.(L.L.)* and the inclusion of police records under the s. 278 scheme together support a finding that the police-healthcare relationship is inextricably linked to the justice system.

2. Compelling Policy Reasons for Class Privilege

Justice L'Heureux-Dubé also drew upon *Gruenke* to conclude that a class privilege should have compelling policy reasons, similar to the solicitor-client privilege.[79] The inextricable link to the justice system described above provided a strong policy

[75] *A.(L.L.) v B.(A.)* (n 21).
[76] ibid [39].
[77] ibid [56]–[60].
[78] ibid [58].
[79] ibid [39].

basis for therapist-patient privilege in *A.(L.L.)* because the privilege would help protect the integrity of the criminal justice system. Pursuant to the discussion above, the same rationale should apply to the police-healthcare privilege. Another policy reason for recognizing privilege in *A.(L.L.)* was that the expectation of confidentiality in the therapist-patient relationship allowed for a 'free flow of discussion which is crucial to the victim's recovery,' which society has an interest in fostering.[80] A similar argument applies to police-healthcare interactions; society has an interest in facilitating communication between these parties to better address the mental health needs of those who come into contact with police.[81]

In addition, a crucial policy argument for a police-healthcare privilege is that improving the protection of complainants' privacy during record disclosure processes is part and parcel of protecting their equality rights. In *A.(L.L.)*, Justice L'Heureux-Dubé acknowledged that the common law principles governing privilege must be consistent with the constitutional values enshrined in the *Charter*.[82] In the context of sexual assault, these values include the complainant's privacy and equality interests. The s. 278 scheme explicitly intends to engage with the *Charter* rights to privacy, equality, and a fair trial. However, as discussed in Section 4, the application of the scheme under *Mills* has resulted in insufficient consideration of equality for women with disabilities. As a result, the policy reasons for a new class privilege can be grounded specifically in the *Charter* equality values.

3. Narrow Class of Actors to Whom Privilege Applies

A.(L.L.) v B.(A.) stressed that class privilege must apply to a category of actors that is limited to specific classes.[83] To this end, Justice L'Heureux-Dubé took issue with the fact that therapeutic relationships cannot be ascribed to a definite class of professionals. Victims of sexual assault may consult with medical professionals as well as unregulated counsellors, community contacts, and friends.[84] On the contrary, privilege between the police and the healthcare system is easily restricted to two types of people: police officers and their medical contacts at psychiatric facilities. The *Mental Health Act* designates and defines the relevant parties in the event of mental health apprehension and could be relied upon to clearly define the scope of the class with respect to health practitioners.[85] The police are already

[80] ibid [56].
[81] Iacobucci (n 52) 111.
[82] *A.(L.L.) v B.(A.)* (n 21) [63].
[83] ibid [70].
[84] ibid [71].
[85] *Mental Health Act* (n 51) s 1.

accepted as a sufficiently narrow class in the context of another privilege, the police-informer privilege.[86]

4. Preserving the Proper Administration of Justice

Despite a history of privileges having to yield in favour of disclosure when a defendant's innocence was in question,[87] recent criticisms of sexual assault law have seriously challenged notions of what information is really necessary for the accused to make a full answer and defence.[88] As Justice L'Heureux-Dubé foresaw in *O'Connor*, it is becoming an accepted view that restricting discriminatory use of complainants' records 'will enhance rather than detract from the fairness of such trials.'[89] Opportunities to 'slag' the complainant covertly based on sexual stereotypes are not necessary in the interests of justice; they are, in fact, forbidden.[90] In *Quesnelle*, the court asserts that not only is it fair for the Crown and police to possess some documents the defence cannot access,[91] but the right to a fair trial is not 'a right to pursue every conceivable tactic to be used in defending oneself against criminal prosecution. The right to a full answer and defence is not without limit.'[92]

Restricting access to communications between the police and healthcare practitioners would not hinder the truth-seeking process because it would prevent discriminatory disclosure, while still allowing opportunity for rightful, relevant disclosure. Essentially, the only unique information that would be entirely protected by privilege and inaccessible elsewhere would be the communications between a police officer and a health professional that are not so formal as to form part of an official report, such as comments, updates and advice on dealing with the mental health challenges of particular individuals. This is not information that was procured in relation to the sexual assault to which the individual is victim. This kind of information would be informal and impressionistic, and would vary greatly in reliability, much like the content of police occurrence reports.[93] Therefore, like highly subjective therapeutic records, this information should generally be treated as having little probative value.[94] To clarify, the police-healthcare privilege would only operate to extent of coordination for purpose of protecting individual and community safety with respect to mental health challenges. Such correspondence would have the primary purpose of serving people in crisis in the community

[86] *R v Scott* [1990] 3 SCR 979, 1990 CanLII 27 (SCC).
[87] *A.(L.L.) v B.(A.)* (n 21) [41].
[88] This was discussed at length in *Mills* and mentioned in *Quesnelle.*
[89] *O'Connor* (n 20) [129].
[90] *Criminal Code* (n 24) s 278.3(4).
[91] *Mills* (n 30) [111].
[92] *Quesnelle* (n 33) [64].
[93] Peter Carmichael Keen, '*Gebrekirstos*: Fallout from *Quesnelle*' (2013), 4 CR (7th) 56, 60–61.
[94] *Mills* (n 30) [136].

as effectively as possible. The privilege would not operate where police are communicating with healthcare professionals for the purpose of investigating an offence. Where there is a genuine issue of credibility or competence to testify, *official* police records and medical records may be sought and disclosed from the police and psychiatric facilities, respectively. This would justifiably exclude informal notes or additional shared information that is not worthy of inclusion in a formal report. Furthermore, given that the privilege would exist between the police and the healthcare institution, if such a privilege were to impede the course of an investigation in any way that would cause injustice, the privilege could be waived by the parties. Therefore, the ability to access relevant records would be maintained.

B. Establishing Statutory Privilege

A privilege for police-healthcare interactions may be better established by statute than at common law. Although the minority decision in *A.(L.L.)* recognised that a new class privilege could be established in theory, Justice L'Heureux-Dubé was hesitant about recognising any new class privilege at common law because this was not a favoured method of protecting privacy; neither historically under Canadian law nor in other commonwealth jurisdictions.[95] Ontario's *Personal Health Information Protection Act (PHIPA)*, the *Police Services Act*, and the *Mental Health Act* present opportunities to further delineate the exchange of mental health information and prescribe measures of privacy protection in a statutory context.

1. Support for Police and Healthcare Confidentiality in Current Legislation

PHIPA already strongly favours confidentiality between healthcare providers and patients; privacy protection of health information is one of its primary functions.[96] *PHIPA* prescribes that health information may be disclosed for the purpose of planning and managing the health system to a prescribed entity with approved privacy and confidentiality measures in place,[97] which might include the police. This function may be expanded for the purpose of managing the mental health system with increased police involvement, pursuant to the TPS report recommendations.

The *Police Services Act* and Ontario Regulation 265/98 also emphasize confidentiality.[98] Disclosure is generally restricted to information about individuals who are being investigated for, have been charged with, or found guilty of an

[95] *A.(L.L.) v B.(A.)* (n 21) [42]–[52].
[96] *Personal Health Information Protection Act*, 2004, SO 2004, c 3, Sch A, s 1(a).
[97] ibid s 45.
[98] *Police Services Act*, RSO 1990, c P.15, s 41(1.1); *Disclosure of Personal Information*, O Reg 265/98.

offence;[99] and is permitted only where it is required for the protection of the public, for the administration of justice, or by law.[100] For agencies not engaged in the former two purposes, such as a hospital, disclosure is made in accordance with a memorandum of understanding between the chief of police and the agency.[101]

The *Mental Health Act* prescribes a more direct relationship between police and hospitals when an individual is apprehended. However, the *Mental Health Act* only limitedly addresses privacy of personal health information, and only from the perspective of the psychiatric facility at that.[102] In any case, the *Mental Health Act* provides a statutory starting point for the privacy mechanisms between police and healthcare facilities to be delineated further.

2. Defending Statutory Privilege Against Charter Challenges

As a statutory creation, the privilege would be subject to *Charter* challenges. Firstly, it would likely see opposition on the basis of encroaching on ss. 7 and 11(d), as the s. 278 scheme did in *Mills*. These arguments would be disposed of on the basis that the privileged information would not actually contribute to a fair trial, and information that might contribute would be available from the police or healthcare facility outside of their privileged communications, as discussed in Section 6.A.4.

The second basis might be that a statutory privilege for personal health information of those who have encounters with police, and not others, is discriminatory. Equality rights are infringed where the legislation's purpose is discriminatory or where it has adverse effects based on an enumerated or analogous ground under s.15 of the *Charter*.[103] It may be argued that provisions creating a privilege for mental health information in possession of police is discriminatory based on mental disability, interpreted broadly to include those who are considered 'emotionally disturbed' in police reports, regardless of the existence or permanence of any medical diagnosis. The privilege might be interpreted as bestowing benefits upon people with mental disability who encounter police, compared to people who encounter police who do not have a mental disability and are not afforded this level of privacy protection.[104]

[99] *Disclosure of Personal Information* (n 97) s 5(1).

[100] ibid s 5(2).

[101] ibid s 5(3).

[102] *Mental Health Act* (n 51) s 35.

[103] *Andrews v Law Society of British Columbia* [1989] 1 SCR 143, 1989 CanLII 2 (SCC); *Withler v Canada (Attorney General)* 2011 SCC 12, [2011] 1 SCR 396.

[104] Interestingly, the police-healthcare privilege would be characterised as a historically more advantaged group, i.e. people without mental disabilities, claiming discriminatory treatment in comparison to a historically disadvantaged group. A similar argument was made in *R v Kapp* 2008 SCC 41, [2008] 2 SCR 483 [*Kapp*], where non-aboriginal fishermen claimed that their s. 15 rights were infringed by legislation that granted aboriginal fishermen the exclusive right to fish one day per year. In that case, the discrimination was characterised as an ameliorative programme under s. 15(2) of the *Charter*.

This argument would be resolved by characterizing the privilege as an ameliorative programme under s. 15(2).[105] The Supreme Court in *Lovelace v Ontario* confirmed that s. 15(2) is an interpretive aid to s. 15(1). Section 15(2) signifies that 'any law, program or activity that has as its object the amelioration of conditions of disadvantaged individuals or groups' is not a form of discrimination within the meaning of s. 15(1), so it does not infringe equality rights.[106] To prove the impugned legislation is an ameliorative programme, the government must show that it has an ameliorative purpose for an identifiable group defined by enumerated or analogous grounds from s. 15.[107] For a police-healthcare privilege, there is clearly an identifiable group: people with mental disabilities who encounter police. An argument for an ameliorative purpose can be based generally on the importance of privilege in promoting a more effective role for police in the mental health system, or it can be based specifically on the failures of the s. 278 application procedure to protect the equality and privacy of women with mental disabilities who have police records.

The court in *R v Kapp* suggests that laws with the purpose of restriction or punishment should not fall under s. 15(2), despite s. 15(2) having been used to uphold criminal laws in lower courts.[108] It is the genuine legislative goal rather than the legislation's actual effects that bring a programme within the scope of s. 15(2).[109] Therefore, the fact that a privilege could indirectly make conviction more likely by removing opportunities to take advantage of mental disability in record disclosure applications, which are not permitted to begin with, is not a constitutional problem; the goal to promote privacy, equality, and a fair trial still fits the meaning of s. 15(2).

7. Conclusion

Equality rights are far from achieving equal status in the *Mills* record disclosure regime. In the context of mental health information in possession of police, sexual myths persist and privacy protection is unpredictable. As long as privacy is presented as a personal right in an adversarial clash with the pursuit of truth, the interests of complainants with mental health issues will likely continue to falter while the role of police in the mental health system expands. A class or statutory privilege, protecting police correspondence with healthcare facilities, would facilitate the creation of a more effective mental health system. It offers an opportunity to bolster equality rights in record disclosure while encouraging a trial that is fair for all persons pursuing justice in sexual assault cases, and by

[105] *Charter* (n 29) s 15(2).
[106] *Lovelace v Ontario* 2000 SCC 37, [2000] 1 SCR 950 [100]–[106].
[107] *Kapp* (n 103) [51]–[55].
[108] ibid [53]–[54].
[109] ibid [46].

all standards of the Canadian criminal law. Just as the record disclosure regime in *O'Connor* transformed into the current *Mills* regime, *Quesnelle*'s affirmation that police records are subject to s. 278 may need to be transformed into a more robust framework for addressing mental health information in possession of police, one that will heighten privacy, equality, and the fairness of proceedings concurrently.

The Past, Present, and Future of Internet Retransmissions of Cable Television: A Suggested FCC Regulatory Framework

Mark DeSantis[1]

1. Introduction

In the twentieth century, broadcast television and the Internet both contributed substantially to the development of American culture and society. Slowly but surely, the two are merging, with the technological benefits of the Internet impacting how and where America views cable television content.[2] Today, viewers no longer watch cable content exclusively by appointment with their televisions.[3] Content is now available on-demand and through streaming services such as Netflix and Hulu.[4] The phenomenon of on-demand and streaming cable content has become extremely popular. In fact, Netflix accounted for 31.6% of all downstream Internet traffic in North America during prime time hours in 2013.[5]

The advent of the Internet has also led some firms to experiment with streaming *live* cable content on devices besides the television, such as laptops,

[1] Juris Doctor, George Mason University School of Law. I would like to thank the editors of the *Cambridge Law Review* for their hard work. I would also like to thank Robert Freedman of Cowan, DeBaets, Abrahams & Sheppard LLP for his valuable insights and edits. Finally, I would like to thank Mack Watson, the author of the article ranked thirteenth, for his crucial contribution.

[2] Michael Marriott, 'Merging TV With the Internet' *New York Times* (New York City, 28 September 2000) <http://www.nytimes.com/2000/09/28/technology/merging-tv-with-the-internet.html> accessed 10 December 2014.

[3] Jesse Cryderman, 'Buyers Guide to Next-Gen Video Platforms' (*Pipeline*, April 2014) <http://www.pipelinepub.com/video_and_content/content_delivery_networks> accessed 10 December 2014.

[4] Adam B. VanWagner, 'Seeking a Clearer Picture: Assessing the Appropriate Regulatory Framework for Broadband Video Distribution' [2011] 79 Fordham L Rev 2909, 2920.

[5] Eliana Dockterman, 'Netflix Used 10 Times More Than Amazon and Hulu Combined' (*Time*, 11 November 2013) <http://entertainment.time.com/2013/11/11/netflix-used-10-times-more-than-amazon-and-hulu-combined> accessed 5 April 2016.

smartphones, and tablets.[6] However, these innovations are proving to be much less successful, mostly because such firms are operating without the permission of cable networks,[7] which would likely have been prohibitively expensive for startup firms to obtain.

Three independent firms—Ivi, Aereo, and FilmOn—have recently encountered legal resistance for streaming live television online.[8] In 2012, several television producers sued Ivi for copyright infringement after the firm began streaming live copyrighted cable content. Ivi argued that it was entitled to a section 111 compulsory licence, which would have allowed it to reproduce the content of cable networks without infringing their copyrights.[9] The Second Circuit disagreed, holding that Ivi did not qualify as a 'cable system' as required by section 111 of the Copyright Act.[10] A similar group of plaintiffs also sued Aereo and FilmOn.[11] Raising a different defence, both firms claimed that they did not 'publicly perform' for purposes of the Copyright Act and thus did not infringe any copyrights.[12] The Supreme Court ultimately rejected this contention, holding that live streaming services do in fact 'publicly perform'.[13] In doing so, however, the Court likened Aereo to a 'cable system'.[14] Aereo viewed this as an overruling of the Second Circuit's *Ivi* decision and subsequently filed for a compulsory licence from the Copyright Office, but the Copyright Office denied their application.[15]

This article will provide a history of cable television, compulsory licences, and Internet retransmissions of cable content before ultimately arguing that section 111 of the Copyright Act could encompass Internet retransmissions if the FCC regulated those retransmissions and required them to be localised. Section 2, Part A will discuss the commercial rise of cable television and the technological struggles that came with it. These technological struggles led to the innovation of community antenna television which was beneficial to programme copyright owners at first but later became detrimental. Part B will explain that Congress provided relief to the programme copyright owners by enacting a compulsory licence for cable retransmissions as part of the 1976 revision of the Copyright Act. Part C will provide a background of the firm Ivi, which was likely the first firm to use the Internet as a medium for cable retransmissions, but was denied a

[6] Dan Garon, 'Poison ivi: Compulsory Licensing and the Future of Internet Television' (2013) 39 Iowa J Corporate L 173, 175.

[7] See *ABC, Inc v Aereo, Inc*, 134 S Ct 2498 (2014); *WPIX, Inc v ivi, Inc*, 691 F 3d 275 (2d Cir 2012); *Fox TV Stations v BarryDriller Content Sys*, 915 F Supp 2d 1138 (CD Cal 2012).

[8] Garon (n 6).

[9] *See ivi* (n 7) [282]; Copyright Act 1976, s 111.

[10] ibid [282].

[11] *See Aereo* (n 7) [2511]; *Fox TV Stations* (n 7) [1146].

[12] ibid.

[13] ibid.

[14] *Aereo* (n 7) [2507].

[15] Keach Hagey, 'Copyright Office Denies Aereo Request to Be Classed as Cable System' *Wall Street Journal* (New York City, 17 July 2014).

compulsory licence. Part D will describe a similar firm, Aereo, which also made secondary retransmissions of cable content on the Internet but did so on the assumption that it did not publicly perform. Part D will also describe Aereo's quest for a compulsory licence.

Finally, Section 3 will analyse Aereo's argument that it was entitled to a section 111 compulsory licence and propose an FCC regulatory framework that would bring Internet retransmissions within the realm of section 111. Part A will analyse the specific merits of Aereo's claims and ultimately reject them. Aereo was not compliant with section 111 because the FCC does not regulate it and because Internet retransmissions are not localised. Part B will then argue that a new FCC regulation could kill two birds with one stone: the FCC could choose to regulate Internet retransmissions and, in doing so, require that those retransmissions be localised utilising geolocation software, solving both of the compatibility issues between Internet retransmissions and section 111.

2. A History of Cable Television, Compulsory Licences, and Streaming Cable Content

Copyright law does not require secondary cable transmitters to negotiate with each copyright holder in the content they transmit.[16] For a statutory fee, the Copyright Office grants them a compulsory licence instead.[17] The concept of secondary transmissions dates back almost as far as cable television itself and an understanding of this history,[18] along with the history of compulsory licences and streaming cable content via the Internet, is necessary to understand why Internet retransmissions are not eligible for a compulsory licence in their current form, and what it would take to make them eligible. Part A will discuss the history of cable television and the technological inadequacies that came with it. Those inadequacies ultimately led to Congress' enactment of section 111, which will be discussed in Part B. Parts C and D will discuss two important pieces of litigation relating to Internet retransmissions of cable content.

A. The History of Broadcast Television and Cable Systems

Cable television first gained commercial success in the 1950s.[19] However, cable signals were originally weak, and many homes received poor reception or no reception at all.[20] Cable owners solved this problem by inventing the community

[16] Copyright Act (n 9) s 111.
[17] ibid.
[18] Garon (n 6).
[19] Fred H Cate, 'Cable Television and the Compulsory License' (1990) 42 Federal Communications LJ 191, 193.
[20] ibid.

antenna television (CATV).[21] Under CATV, a community shared one large antenna, which picked up signals more clearly from local television broadcast stations, and individual users would connect to this large antenna with a coaxial cable.[22] Initially, both television broadcasters and copyright owners encouraged this practice; an increased viewer base meant higher advertising prices and royalty fees.[23] However, technology eventually improved, and with it, the distances the signals could travel increased. As a result, broadcast networks lost their local market monopolies.[24]

The advent of colour television worsened matters. Signal interference resulting from tall buildings in urban areas hardly affected black and white television, but the effect was noticeable with colour television.[25] In addition, cable operators began broadcasting their own programmes which were not available on broadcast television.[26] Consequently, as viewers began making the switch to cable television copyright owners began losing compensation because the cable operators did not pay them a royalty or licence fee.[27] As a result, copyright owners and the broadcast industry began suing cable systems for copyright infringement, alleging that the retransmissions were 'performances' for purposes of the Copyright Act 1909.[28] The Supreme Court disagreed with this contention in *Fortnightly Corp v United Artists Television, Inc* in 1968, ruling in favour of the cable systems.[29] Six years later, the Supreme Court again held that broadcast retransmissions did not infringe copyrights in *Teleprompter Corp v Columbia Broad Sys, Inc*.[30] The copyright owners and the broadcast industry had to seek relief elsewhere.

[21] ibid.
[22] ibid.
[23] Garon (n 6) 180.
[24] ibid.
[25] Cate (n 20) 195.
[26] ibid.
[27] ibid.
[28] See generally *Fortnightly Corp v United Artists Television, Inc*, 392 US 390, 400–402 (1968); *Teleprompter Corp v Columbia Broad Sys, Inc* 415 US 394, 409 (1974). The right to publicly perform is an exclusive right granted to copyright owners. If the CATVs 'performed' the copyrighted cable content, they were committing copyright infringement. See Copyright Act (n 9) s 106(4), 501(a).
[29] *Fortnightly* (n 28) [400]–[402].
[30] *Teleprompter* (n 28) [409].

B. Congress Enacts a Compulsory Licence

Congress responded swiftly to *Fortnightly* and *Teleprompter* in its 1976 revision of the Copyright Act.[31] While the statute already defined 'public performance', Congress added to it as follows:

> To perform or display a work 'publicly' means…to *transmit* or otherwise communicate a performance or display of the work... to the public, by means of any device or process, whether the members of the public capable of receiving the performance or display receive it in the same place or in separate places and at the same time or at different times.[32]

These changes directly brought cable retransmissions within the definition of a 'public performance', thus effectively overruling *Fortnightly* and *Teleprompter*. But Congress did not stop there: the 1976 revision to the Copyright Act also created a compulsory licence for the same cable retransmissions they brought into the realm of 'public performances'.[33] A compulsory licence requires cable operators to pay a statutory royalty to the Copyright Office, as opposed to negotiating a royalty with each copyright holder.[34] The Copyright Office then disburses the royalties to the copyright holders.[35] This compulsory licence applies exclusively to retransmissions that meet the definition of a 'cable system'. The Copyright Act defines a 'cable system' as a facility that 'receives signals transmitted or programmes broadcast by one or more television broadcast stations licensed by the Federal Communications Commission, and makes secondary transmissions of such signals or programs by wires, cables, microwave, or other communications channels to subscribing members of the public who pay for such service.'[36]

The Act also specifies requirements with which each cable system must comply. Cable systems must make an annual 'Statement of Account' setting out the number of rebroadcasts they make.[37] They also may not modify the transmissions in any way.[38] Finally, the cable systems must adhere to all FCC regulations.[39] To date, the Copyright Act has added two additional compulsory licences, both of which are for satellite carriers.[40] The two additional licences were a response to

[31] Copyright Act (n 9) s 101.
[32] ibid (emphasis added).
[33] ibid.
[34] Copyright Act (n 9) s 111.
[35] ibid.
[36] ibid.
[37] ibid.
[38] ibid.
[39] ibid.
[40] Copyright Act (n 9) ss 119, 122.

the development of a new medium, satellite, for viewing cable content. However, Congress has yet to respond to a more recent development of a cable viewing medium: the Internet. As the Internet grew in popularity, so too did the demand for viewing cable content via the Internet. One of the first firms to stream live cable content over the Internet was called Ivi.

C. WPIX, Inc v IVI, Inc

On September 13, 2010, a firm called Ivi released its 'revolutionary live television application', which it claimed 'enable[s] anyone with an Internet connection' to 'watch live television anywhere in the world, anytime.'[41] The launch was the first of its kind.[42] By downloading an app on Ivi's website, users could start with a free 30 day trial and stream the broadcasts of major networks such as ABC, NBC, CBS, and Fox.[43] Essentially, Ivi's premise was that it brought television to the Internet while other firms focused on bringing the Internet to television.[44] Ivi also allowed its users to 'cut the cord'.[45] Indeed, Ivi brought users a significant advantage over traditional television: users could watch anything a network's affiliates in New York, Los Angeles, Chicago, or Seattle were currently streaming.[46] Traditional television limits viewers to watching only broadcasts from local stations.[47] Ivi did not obtain the consent of any of the broadcasters it streamed.[48]

The broadcasters' response was swift. They sent several cease and desist letters to Ivi, to which Ivi was not responsive.[49] About two weeks after Ivi's launch, the broadcasters brought suit in federal court in the Southern District of New York seeking, among other things, a preliminary injunction.[50] Distributors of non-commercial education programmes, Major League Baseball, top motion picture studios and individual broadcast television stations joined the broadcasters as plaintiffs.[51] The plaintiffs contended that Ivi's streaming service was an unsanctioned 'public performance' of their copyrighted works.[52] Section 106 of the Copyright Act gives a copyright owner several exclusive rights, including

[41] Hal Bringman, 'ivi, Inc. Launches Highly Disruptive Software Delivering Live TV to the Internet' (*PRWeb*, 13 September 2010) <http://www.prweb.com/releases/2010/09/prweb4487284.htm> accessed 10 December 2015.

[42] ibid.

[43] ibid.

[44] ibid.

[45] 'Cutting the cord' refers to the concept of ceasing one's cable subscription in favour of using the Internet to view cable content.

[46] *WPIX, Inc v ivi, Inc*, 765 F Supp 2d 594, 599 (SDNY 2011) affd 691 F 3d 275 (2d Cir 2012).

[47] ibid.

[48] ibid.

[49] ibid.

[50] ibid.

[51] ibid.

[52] ibid.

the right to publicly perform their work.[53] The violation of any of the section 106 exclusive rights constitutes copyright infringement.[54] Ivi argued that it fit within the statutory definition of 'cable system' as provided by section 111 of the Copyright Act.[55] It made this argument by pointing out that its facility was located in the United States and received signals transmitted by broadcast stations for a secondary transmission.[56] While conceding that they did not comply with the 'rules, regulations, or authorizations of the Federal Communications Commission' as required by section 111 of the Copyright Act,[57] Ivi contended that its transmissions were permissible because the FCC does not, in fact, regulate the Internet.[58]

The District Court disagreed, noting that no technology had ever been allowed to take advantage of section 111's compulsory licence without complying with the rules and regulations of the FCC.[59] The Court buttressed its conclusion by looking to the legislative history of section 111 and taking into account practical considerations. Specifically, the Court found it significant that Congress understood the cable system to be a 'highly localized medium', which the Internet is not, and that Ivi refused to comply with the rules and regulations of the FCC.[60] In addition, the Court granted Skidmore deference to the Copyright Office's interpretation of section 111[61] and reviewed several pieces of evidence that suggested that the Copyright Office disapproved of granting compulsory licences for Internet retransmissions.[62] Under Skidmore deference, a Court gives weight to an agency's determinations to the extent that the agency's judgment is persuasive.[63] The Copyright Office, like the Court, found the reach of an Internet retransmission to be too broad, and not 'localized' in the sense that Congress intended.[64] Ivi made one final argument: they should be included as a 'cable system' under section 111 due to the broadness of the statute's definition of a 'cable system'.[65] Once

[53] Copyright Act (n 9) s 106.

[54] ibid (n 9) s 501(a).

[55] *ivi* (lower court) (n 46) [599].

[56] Memorandum in Support of Motion to Stay Preliminary Injunction Pending Appeal, *WPIX, Inc v ivi, Inc*, 765 F Supp 2d 594 (SDNY 2011) (Docket No 10 Civ 7415 (NRB)).

[57] Copyright Act (n 9) s 111(b)(2); *ivi* (lower court) (n 46) [599].

[58] *ivi* (lower court) (n 46) [599].

[59] ibid [602]. This held that Ivi is not a cable system under section 111 of the Copyright Act.

[60] ibid [604].

[61] ibid [605].

[62] ibid [609]–[610] (quoting US Copyright Office, *A Review of the Copyright Licensing Regimes Covering Retransmission of Broadcast Signals* 97 (1997)).

[63] *Skidmore v Swift & Co*, 323 US 134, 137 (1944).

[64] *ivi* (lower court) (n 46) [609]–[610] (quoting US Copyright Office, *A Review of the Copyright Licensing Regimes Covering Retransmission of Broadcast Signals* 97 (1997)).

[65] ibid [616]. Ivi relied solely on the first part of the definition, which states that a cable system is a 'facility, located in any state…that in whole or in part receives signals transmitted or programs broadcast...and makes secondary transmissions of such signals or programs by wires, cables, microwave, or other communications for channels to subscribing members of the public who pay for such service' (quoting Copyright Act (n 8) s 111(f)(3)).

again, the Court was not persuaded. It noted that Ivi's interpretation neglected the second sentence of section 111(f)(3), which refers to 'headends' and 'contiguous communities', two concepts not present in Ivi's technology.[66] Accordingly, the Court held that Ivi was not likely to succeed on the merits of its case and ultimately granted the plaintiffs a preliminary injunction.[67]

Shortly thereafter, Ivi appealed to the Court of Appeals for the Second Circuit.[68] In deciding the issue of a preliminary injunction and the plaintiffs' likelihood of success on the merits of their case, the Court of Appeals focused solely on the Copyright Office's interpretation of section 111.[69] Unlike the District Court, the Court of Appeals gave the Copyright Office Chevron deference, which is stronger than the Skidmore deference the District Court granted.[70] At Chevron step one, the Court inquires whether Congress directly spoke to the issue at hand.[71] If Congress did not directly speak to the issue, the Court proceeds to step two, asking whether an agency's interpretation is 'permissible'.[72] At Chevron step one, the Court determined that the statutory text of section 111 was ambiguous as to whether an Internet retransmission is eligible for a compulsory licence and accordingly proceeded to step two.[73] The 'thoroughness' and 'validity' of the Copyright Office's reasoning in interpreting section 111 was sufficient for the Court.[74] The Court also held that the legislative history of section 111 revealed that Congress did not intend for Internet retransmissions to be eligible for a compulsory licence.[75] For those reasons, the Court deemed the Copyright Office's interpretation that section 111's definition of 'cable system' did not encompass internet providers to be permissible.[76] Ivi lost its legal battle, but it was not the last Internet company to fight for the right to retransmit live streaming cable online.

[66] ibid [616]. A headend is the point at which cable signals are monitored and processed before being distributed. Contiguous communities are neighboring areas controlled by one headend.

[67] ibid [617]–[622].

[68] *Ivi* (n 7) [278].

[69] *See generally* ibid.

[70] ibid (n 7) [279].

[71] *Chevron USA, Inc v Natural Resources Defense Council, Inc*, 467 US 837 (1984).

[72] ibid.

[73] *ivi* (n 7) [280]. The Court of Appeals thought 'facility' was ambiguous and doubted that the Internet qualified as such.

[74] The weight a court gives to an agency's interpretation of a statute 'depend[s] upon the thoroughness evident in its consideration, the validity of its reasoning, its consistency with earlier and later pronouncements, and all those factors which give it power to persuade.' *United States v Mead Corp*, 533 US 218, 228 (quoting *Skidmore v Swift & Co*, 323 US 134, 140 (1944)).

[75] *ivi* (n 7) [280].

[76] ibid [284]–[285] (citing *Chevron* [837]; Copyright Act (n 9) s 111).

D. ABC, Inc v Aereo, Inc

Like Ivi, a firm called Aereo, Inc ('Aereo') also tried its hand at streaming the live broadcasts of copyrighted television programmes.[77] Aereo operated differently: for each of its users, it allocated one antenna the size of a dime and one transcoder that in turn transmitted copyrighted content at the user's request.[78] When a viewer chose content to stream, an antenna server operated by Aereo sent a 'tune request' directing the viewer's antenna to tune into a specified broadband frequency correlated with the desired broadcast.[79] In addition, Aereo gave its viewers the option to watch *or record* the content they streamed.[80] When a viewer watched content, they also had the option to pause or rewind the content.[81] In this respect, Aereo was similar to the digital video recorders (DVR) that cable providers offered, except that it operated via computers, laptops and mobile devices.[82] When a viewer chose to record the content, Aereo saved the stream to a permanent hard disk the viewer could later access, which it did not do when the viewer only chose to watch the content.[83]

On March 1, 2012, a similar group of copyright holders as those in *Ivi* brought a claim against Aereo in the District Court for the Southern District of New York and moved for an injunction to prevent Aereo from allowing its users to stream live broadcasts.[84] Relying on a Second Circuit case, *Cartoon Network, LP v CSC Holdings, Inc* (*Cablevision*), Aereo contended that it was not in violation of copyright law.[85] *Cablevision* involved a dispute over an RS-DVR system that allowed users to record cable programming on a hard drive system that Cablevision operated at a remote location, much as Aereo operated its recording mechanism at a remote location.[86] In *Cablevision*, the Second Circuit held that the video streams of DVRs were not 'public performances' for purposes of the Copyright Act's Transmit Clause because only one person received each transmission.[87] The District Court in the Aereo litigation concluded that the two cases were indistinguishable, and thus denied the injunction.[88]

[77] *ABC v Aereo, Inc*, 874 F Supp 2d 373, 377 (SDNY 2012); revd *ABC, Inc v Aereo, Inc*, 134 S Ct 2498 (2014).

[78] ibid [379]. Unlike the CATV system, which led to the creation of statutory licenses, Aereo users did not share a satellite; one was allocated to each user.

[79] ibid [378].

[80] ibid [377].

[81] ibid.

[82] ibid.

[83] ibid.

[84] ibid [376]; Garon (n 6) 192.

[85] *Aereo* (n 77) [373].

[86] ibid (citing *Cablevision* [124]).

[87] *Cablevision* [137].

[88] *Aereo* (n 77) [405].

The plaintiffs appealed to the Court of Appeals for the Second Circuit on November 30, 2012, which reviewed the District Court's denial of a preliminary injunction for abuse of discretion.[89] The Second Circuit affirmed the lower court's holding. The Court first interpreted *Cablevision* and identified four guideposts relevant to Aereo.[90] First, if the public is capable of receiving a transmission, that transmission is a public performance.[91] However, if only one person is capable of receiving a transmission, it is not a public performance.[92] The second guidepost was a corollary of the first: courts cannot aggregate private transmissions and call them public performances, except as provided for in the third guidepost.[93] According to the third guidepost, courts should aggregate private transmissions when private transmissions all result from the same copy of the work.[94] If the aggregated transmissions from that single copy enable public viewing, that transmission is a *public* performance.[95] Finally, courts should give weight to factors that limit a potential audience for the purposes of the Transmit Clause. With these guideposts in place, the Second Circuit applied *Cablevision* to the facts before them.[96] It found the two cases to be indistinguishable: the RS-DVR system in *Cablevision* created unique copies of the programme a user wished to record, and the transmission was also generated from that unique copy.[97] Aereo's transmissions were unique copies transmitted at the user's request while the programmes were still on broadcast television.[98] The Second Circuit held Aereo's streaming did not constitute a public performance and upheld the District Court's denial of a preliminary injunction.[99]

The Supreme Court granted certiorari in the case to address two questions: whether Aereo 'performed' at all and, if so, whether Aereo performed publicly.[100] The Court answered the first question in the affirmative, reasoning that both Aereo and the viewer of a television programme 'performed' when they used Aereo's streaming service.[101] The Court next turned to the issue of whether Aereo's performance was public.[102] Aereo argued that the performance was not public, since each antenna was allocated to just one subscriber and thus only one subscriber had the ability to view each transmission.[103] The Supreme Court was

[89] *WNET v Aereo, Inc*, 712 F 3d 676 (2d Cir 2012); revd *ABC, Inc v Aereo, Inc*, 134 S Ct 2498 (2014).
[90] ibid.
[91] ibid [689].
[92] ibid.
[93] ibid.
[94] ibid.
[95] ibid.
[96] ibid [689]–[694].
[97] ibid (citing *Cablevision* [124]).
[98] ibid [696].
[99] ibid (citing *Cablevision* [124]).
[100] *ABC, Inc v Aereo, Inc* (n 89) [2504].
[101] ibid [2506]–[2508].
[102] ibid [2509].
[103] ibid [2508].

not convinced. In terms of Congress' regulatory objectives, the technological differences between Aereo and a cable provider were irrelevant.[104] These 'behind-the-scenes' mechanics did not change Aereo's commercial objective.[105] The Court concluded that Aereo did indeed perform publicly.[106] Justice Stephen Breyer's opinion analogised Aereo to a cable system and he considered Aereo's practice highly similar to the CATV systems in *Fortnightly* and *Teleprompter*.[107] Accordingly, the Supreme Court reversed the Second Circuit's decision and ruled that Aereo was infringing numerous copyrights.[108]

Aereo may not have gotten the ruling it wanted from the Supreme Court, but it was pleased with the Supreme Court's analogy of its service to a cable system.[109] Since cable systems are generally entitled to a compulsory licence,[110] Aereo sent a cheque for $5,310.74 to the Copyright Office in hopes of obtaining a compulsory licence which would make the Supreme Court's ruling that it 'publicly perform[ed]' irrelevant.[111] The Copyright Office, however, did not accept Aereo's money.[112] They cited *Ivi* and refused to grant Aereo a compulsory licence.[113] Dissatisfied with the Copyright Office's ruling, Aereo decided to attempt litigation one more time and brought suit in the District Court for the Southern District of New York, contending that the plaintiffs from its recent Supreme Court case were not entitled to a preliminary injunction despite their victory.[114] This contention was based on the new affirmative defence that Aereo was entitled to a compulsory licence because Justice Breyer's opinion essentially overruled the *Fortnightly* and *Teleprompter* decisions.[115] The Court disagreed, reasoning that Aereo's similarity to a cable system did not necessarily entitle it to the compulsory licence granted to genuine cable systems under section 111.[116] In addition, the Supreme Court never addressed the issue of compulsory licences when deciding *Aereo*, a void which *Ivi* seemingly filled.[117] For those reasons, the Court granted the plaintiffs an injunction.[118]

[104] ibid.
[105] ibid [2509].
[106] ibid [2511].
[107] ibid.
[108] ibid.
[109] Hagey (n 15).
[110] Copyright Act (n 9) s 111.
[111] ibid; Hagey (n 15).
[112] ibid.
[113] Letter from Jacqueline C Charlesworth, General Counsel and Associate Register of Copyrights, to Matthew Calabro, Aereo, Inc (July 16, 2014).
[114] *ABC v Aereo, Inc,* 2014 US Dist LEXIS 150555 (SDNY Oct 23, 2014).
[115] ibid.
[116] ibid.
[117] ibid.
[118] ibid.

Aereo's fight was seemingly still not over after losing in the District Court on remand. On October 28, 2014, FCC Chairman Tom Wheeler posted to the Official FCC Blog that the FCC would consider enacting rules to regulate Internet retransmissions such as Aereo.[119] According to Wheeler's post, the FCC wants to open access to cable programmes for Internet video services.[120] Wheeler compared the up-and-coming technology to that of satellites when satellites first transmitted television content.[121] He felt that new rules could spur competition and that the Internet as a medium could allow consumers to purchase smaller cable packages.[122] Judging from Wheeler's post, the FCC appears enthusiastic about enacting new rules, and their creation may be imminent.

Ultimately, Aereo was unable to overcome its legal issues. It filed for bankruptcy on November 20, 2014.[123] Its Chapter 11 filing marked the end of the legal and financial troubles that had defined the company for the past year.[124] New FCC rules may be too late for Aereo, but they could potentially pave the way for similar firms to gain access to a section 111 compulsory licence. It is therefore essential to examine whether future Internet retransmissions of cable content could ever be eligible for a section 111 compulsory licence.

3. Aereo's Claims Under Section 111 and the Future for Internet Retransmissions

Although Aereo is now bankrupt, it is worth analysing the merits of their section 111 claims since those claims were decided by a District Court and the FCC is interested in enacting rules to cater to similar services. In addition, it is likely that Aereo will not be the last firm to experiment with Internet retransmissions of cable content. For example, FilmOn continues to operate and it provides a similar service to that of Aereo.[125] Part A of this Section will analyse the merits of Aereo's claims under section 111 and conclude that the District Court rightfully disposed of those claims. Aereo did not comply with section 111 because it did not take certain steps *before* commencing its retransmitting service. Part B will examine whether

[119] Tom Wheeler, 'Tech Transitions, Video, and the Future' (*Official FCC Blog*, 28 October 2014), <http://www.fcc.gov/blog/tech-transitions-video-and-future> accessed 10 December 2014.

[120] ibid.

[121] ibid.

[122] ibid.

[123] Tanya Agrawal and Jonathan Stempel, 'Video streaming service Aereo files for bankruptcy' (*Reuters*, 21 November 2014) <http://www.reuters.com/article/2014/11/21/us-aereo-bankruptcy-idUSKCN0J513K20141121> accessed 10 December 2014.

[124] Emily Steel, 'Aereo Concedes Defeat and Files for Bankruptcy' *The New York Times* (New York City, 21 November 2014).

[125] Lisa Shuchman, 'FilmOn Fights On After Aereo Bankruptcy' (*The AM Law Litigation Daily*, 24 November 2014) <http://www.litigationdaily.com/home/id=1202677363695?mcode=1202617031029&curindex=0&slreturn=20150009090918> accessed 10 December 2014.

an Internet service similar to Aereo could be eligible for a section 111 licence if it were to take those steps before commencing operations. Because the Internet is not localised or regulated by the FCC, Part B will conclude that other Internet retransmissions are also ineligible. Part C will suggest a regulatory framework for the FCC to adopt that would bring Internet retransmissions within the realm of section 111.

A. Analysing the Merits of Aereo's defence on Remand from the Supreme Court's Decision

On remand from the Supreme Court decision, the District Court properly disposed of Aereo's claim that it was entitled to a compulsory licence because the FCC did not regulate Aereo and Aereo was not compliant with the plain requirements of section 111. At first blush, it may appear that Aereo was a 'cable system' as contemplated by Congress when it passed the Copyright Act of 1976. The corresponding Senate Report states that cable systems 'are commercial subscription services that pick up broadcasts of programmes originated by others and retransmit them to paying subscribers'.[126] Aereo seemingly matched this description; it used antennas to retransmit the broadcasts of others and relayed them to paying subscribers.[127]

However, a closer look at the wording of section 111 reveals that Aereo did not comply with the plain requirements for obtaining a compulsory licence. Aereo did not apply for a compulsory licence until Justice Breyer analogised them to a 'cable system' in his *Aereo* opinion.[128] Under section 111, cable systems must take specific steps one month before commencing operations to be eligible for a compulsory licence.[129] Specifically, Aereo needed to record a notice in the Copyright Office including its identity and address along with the name and location of the primary transmitter whose signals it regularly carries.[130] There is no indication that Aereo took these steps, likely because it designed its business model to take advantage of the Second Circuit *Cablevision* decision, which concerned the public performance right.[131] Aereo created its service in hopes that it did not infringe copyrights at all, and therefore did not need a compulsory licence in the first place.

Even without the clarity of section 111, courts grant the Copyright Office Chevron deference in their interpretations of the Copyright Act.[132] Although Congress may not have directly spoken to the issue of Internet retransmissions,

[126] S Rep 94-473 (1975).

[127] *ABC v Aereo, Inc*, 874 F Supp 2d 373, 378 (SDNY 2012); revd *ABC, Inc v Aereo, Inc*, 134 S Ct 2498 (2014).

[128] *See* n 109–111.

[129] Copyright Act (n 9) s 111(d).

[130] HR Rep No 1476, 94th Cong, 2d Sess 95 (1976).

[131] Timothy B Lee, 'With Aereo appeal, broadcasters threaten the foundation of locker services' (*Ars Technica*, 17 April 2013) <http://arstechnica.com/tech-policy/2013/04/with-aereo-appeal-broadcasters-threaten-the-foundation-of-locker-services/> accessed 10 December 2014.

[132] *See ivi* (n 7) [280].

Courts still uphold an agency's interpretation in such situations as long as the interpretation is 'permissible'.[133] The Copyright Office's interpretation easily meets this relatively low standard. It denied Aereo's application because Aereo's service was not localised and the FCC did not regulate it.[134] The wording and legislative history of section 111 so clearly beg this determination that, arguably, an *opposite* finding would not be 'permissible'. Therefore, the District Court for the Southern District of New York properly disposed of Aereo's affirmative defence that it was entitled to a compulsory licence on remand from the Supreme Court's decision. But does this mean that the future of Internet retransmissions of cable content is doomed?

B. Are Other Internet Retransmissions of Cable Content Necessarily Ineligible for a Section 111 Compulsory Licence?

Other Internet retransmissions, besides Aereo, face similar difficulties in their applicability to section 111. Internet retransmissions can hardly be classified as 'localized' because the Internet's reach is extremely vast. Content on the Internet is not just available across the nation, but internationally as well.[135] As the Courts in *Ivi* and *Aereo* correctly held, Congress intended section 111 of the Copyright Act to apply to *localised* retransmissions only.[136] This much is plainly clear from the legislative history of section 111.[137] Congress' entire purpose in enacting section 111 was to provide copyright holders relief while maintaining the conveniences and benefits of the local secondary cable transmissions already in place.[138] Specifically, Congress was concerned with preserving the then-current economic state of cable retransmissions.[139] The Committee on the Judiciary in the House of Representatives believed that the retransmission of 'local' broadcast signals posed no threat to the existing market for copyright holders in cable content.[140] The Committee also found it significant that networks compensated those copyright holders based on the local markets the networks served.[141] As a corollary, it believed that transmission of distant programming would 'adversely affect the ability of

[133] *Chevron* (n 71) [843].

[134] Letter from Jacqueline C Charlesworth, General Counsel and Associate Register of Copyrights, to Matthew Calabro, Aereo, Inc (16 July 2014).

[135] Dan Jerker B Svantesson, 'Geo-Location Technologies and Other Means of Placing Borders on the "Borderless" Internet' (2004) 23 John Marshall J Computer & Information L 101.

[136] *Ivi* (n 7) [280]; *Aereo* (n 114); '…the Committee has concluded that the copyright liability of cable television systems under the compulsory license should be limited to the retransmission of *distant* nonnetwork programming', HR Rep No 1476, 94th Cong, 2d Sess 99 (1976) (emphasis added).

[137] 57 Fed Reg 3284 (29 January 1992).

[138] HR Rep No 1476, 94th Cong, 2d Sess 99 (1976) (noting that 'distant' signals are not to be subject to payment under the section 111 licence); Cate (n 19) 195.

[139] HR Rep No 1476, 94th Cong, 2d Sess 90 (1976).

[140] ibid.

[141] ibid.

the copyright owner to exploit the work in the distant market.'[142] Since Congress intended to preserve the then-current economic state of secondary retransmissions, any interpretation of section 111 that disrupted that economic state would be contrary to Congress' will. Because of Congress' stated intention, it is proper to consider the economic ramifications of permitting Internet retransmissions to obtain a compulsory licence under section 111.

1. The Economics of Allowing Section 111 in its Current Form to Cover Internet Retransmissions

If the phenomenon of independent firms retransmitting cable signals nationwide becomes widespread, which it would if anyone was able to do so by simply applying for a compulsory licence, the business models of many cable providers would collapse. The cable industry depends on the long-standing practice of marketing television programmes on a geographic basis.[143] Specifically, advertising is made specially for localised transmissions.[144] Although television series are made for a national, or even international, audience, the series and other cable content are subject to numerous retransmissions throughout the United States.[145] Accordingly, the market is segmented.[146] The localised nature of a traditional cable retransmission adds value to the cable content because it allows advertisers to target local audiences and provides opportunities for local businesses to narrowly advertise within their locales. But the segmented market resulting from local retransmissions is not the only economic concern.

The large quantity of local cable retransmissions has also created a reliance interest in the cable industry. Because traditional cable retransmission systems only reach a limited distance, there must be many of them to cover the vast number of regions in the United States. This fact highlights the need for a compulsory licence in the first place: the retransmissions are so numerous that it would be financially prohibitive to require each retransmitter to negotiate with the holder of each copyright they 'perform'.[147] The Copyright Royalty Board sets the statutory fee for the compulsory licence based on the large quantity of licences they will grant and their market value, as determined in part by advertising revenue.[148] Clearly, the cable industry has a legitimate reliance interest in the cable retransmission system in its current form because of the advertising business model

[142] ibid.
[143] Matt Jackson, 'The Technological Revolution Will Not Be Televised: Canadian Copyright and Internet Retransmissions' (2006) Michigan State L Rev 133.
[144] ibid.
[145] ibid.
[146] ibid.
[147] *Ivi* (n 7) [278].
[148] Copyright Act (n 9) s 111(d)(A).

that has been in place for years. The 'borderless' nature of the Internet would upset this balance already in place and wreak havoc on the cable industry.[149]

Accordingly, economic considerations, in addition to the legal considerations addressed above, suggest that secondary retransmissions over the unrestricted Internet are ineligible for a section 111 compulsory licence. But this is not to say that the Internet itself can never be a suitable medium for viewing cable content. If the FCC adopted a framework for regulating Internet retransmissions that required those retransmissions to be localised, Internet retransmissions could potentially be eligible for a section 111 compulsory licence.

C. Killing Two Birds with One Stone: A Proposed Regulatory Framework for the FCC to Apply to Internet Retransmissions of Cable Content

On October 28, 2014, after Aereo lost both its Supreme Court challenge and its challenge to obtain a compulsory licence, the Chairman of the FCC announced that the FCC would begin the process of changing its rules to accommodate Aereo and thus allow Aereo to operate as a cable system.[150] Although Aereo is now bankrupt, new FCC rules could apply to future firms launching a similar service. In fact, FilmOn, a firm that competed with Aereo and also streams live cable content, is still in business, and such a rule change would pave the way for FilmOn and similar firms to obtain compulsory licences.[151]

The FCC would not be overstepping its bounds by enacting such rules. Congress clearly reserved a role for the FCC in the compulsory licence scheme by requiring cable systems to comply with FCC rules and regulations.[152] In addition, the legislative history of section 111 reveals that Congress did not intend for section 111 to affect communications policy, which they desired to be the FCC's prerogative.[153] In other words, Congress recognised the interplay between communications policy and copyright law and intended to affect only the latter when it enacted section 111. With a properly crafted regulatory framework, the FCC could do its part in the interplay between communications policy and

[149] Jackson (n 143).

[150] Joshua Brustein, 'The FCC Wants to Let Aereo Become a Cable Service' (*Bloomberg Businessweek*, 28 October 2014) <http://www.businessweek.com/articles/2014-10-28/the-fcc-wants-to-let-aereo-become-a-cable-network> accessed 10 December 2014.

[151] Shuchman (n 125); Shalini Ramachandran, 'Aereo Investors See a "Plan B" After FCC's Latest Move' (*CMOToday*, 31 October 2014) <http://blogs.wsj.com/cmo/2014/10/31/aereo-investors-see-a-plan-b-after-fccs-latest-move/> accessed 10 December 2014.

[152] Copyright Act (n 9) s 111(a)(1).

[153] 'While the Committee has carefully avoided including in the bill any provisions which would interfere with the FCC's rules or which might be characterised as affecting "communications policy", the Committee has been cognizant of the interplay between the copyright and the communications elements of the legislation.' HR Rep No 1476, 94th Cong, 2d Sess 89 (1976).

copyright law and bring Internet retransmissions within the realm of section 111, enabling the previously elusive marriage of live cable and the Internet.

First, an FCC rule change recognizing Internet retransmissions as cable systems would itself solve one of the compliance issues with the text of section 111. The statutory text was previously a roadblock because Aereo was not compliant with FCC rules and regulations as required by section 111(a)(1).[154] As it turns out, there were no rules and regulations to follow because the FCC does not regulate Internet retransmissions of cable content at all. The FCC now desires to treat Internet retransmissions as cable systems and thus could bring them under its rules. New FCC rules, regardless of their content, would bring Internet retransmissions into compliance with FCC regulations, as required by section 111.[155] Second, the FCC could kill two birds with one stone by mandating that Internet retransmissions be localised, requiring retransmission services to adopt specific technology to address the problem of the Internet's global nature.[156] Such technology already exists and is known as geolocation software.[157] Geolocation software identifies the location of an Internet user and can report the location to a website that retransmits cable content.[158] The FCC could require websites to authenticate the location of a user using geolocation software to ensure that the viewer is truly local. As a result, Internet retransmissions would serve localised markets in the same manner as traditional cable secondary retransmissions, thereby bringing Internet retransmissions within the scope of section 111 and benefitting society through the expanded interchange of information and ideas—all without upending the well-established local balance of the cable industry.

4. Conclusion

Secondary retransmission of cable television content is almost as old as cable television itself. Retransmission was originally a solution to the poor signal quality that consumers experienced when cable television first became commercially successful. The retransmission process segmented the market for cable television as smaller communities received their own transmissions. This market segmentation allowed advertisers to focus on different locales, giving rise to the economic reality of cable television as it exists today. When Congress enacted the compulsory licence provision in section 111 of the Copyright Act, it intended to preserve this status quo, enabling 'cable systems' to receive compulsory licences only if they were

[154] Copyright Act (n 9) s 111(a)(1).

[155] ibid.

[156] Congress enacted section 111 to provide compulsory licences only for *localised* secondary retransmissions of cable. *See* n 138–142 and accompanying text.

[157] Daniel Ionescu, 'Geolocation 101: How It Works, the Apps, and Your Privacy' (TechHive, 29 March 2010) <http://www.techhive.com/article/192803/geolo.html> accessed 10 December 2014.

[158] ibid.

regulated by the FCC and their retransmissions were localised. Because Congress passed this licensing scheme in 1976, it likely did not consider the possibility of Internet retransmissions.

Ivi was the first firm to retransmit cable content over the Internet, but the Second Circuit put an end to this venture, ruling that Internet retransmissions of cable content were ineligible for a compulsory licence. Aereo came on the scene shortly thereafter, offering a service highly similar to that of Ivi, but purporting not to infringe copyrights because it did not 'publicly perform'. Its fate, however, was similar to Ivi's, leaving the future of Internet retransmissions of cable content uncertain. Aereo was not entitled to a compulsory licence because it was not regulated by the FCC and because its retransmissions were not localised. The FCC, however, could kill both birds with one stone by stepping in to regulate Internet retransmissions. Regardless of the content of its regulations, by adopting a regulatory scheme for Internet retransmissions, the FCC would remove the first barrier to compulsory licenses for services like Aereo. The FCC could remove the second barrier—the localization problem—by requiring Internet retransmissions to utilise geolocation software, thereby ensuring that users are truly 'local'. This requirement would satisfy Congress' intent that only localised transmissions be protected under section 111. It would also maintain the segmented nature of the cable television market that Congress sought to preserve. In sum, an FCC rulemaking on Internet retransmissions could bring cable content into the twenty-first century.

Mosh Pits or Liability Pits: Criminal and Tortious Liability at Concerts

Thomas Charles Surmanski[1]

Our enemies have beat us to the pit.
It is more worthy to leap in ourselves,
Than tarry till they push us.
(*Julius Caesar* 5.5.27-29)

1. Introduction

In an American case, a mosh pit patron, Kimberly Myers, was assaulted by the band itself.[2] Myers attended a Fishbone ska-punk concert in 2010. During the concert, Fishbone's lead singer Angelo 'Dr. Madd Vibe' Moore, dove from the stage crushing her. She suffered a broken skull and collarbone.[3] Following the incident, and Myers losing consciousness, Fishbone 'continued to perform as if nothing had happened.'[4] The defendant showed no real remorse for the incident and stated that he gives no warning before stage diving as it would ruin the 'theatrics' of his performance.[5] Moore added that '[p]eople want to be on the edge when they go to a Fishbone show.'[6] U.S. District Judge Jan DuBois rejected this excuse and ordered Moore and bassist John Norwood to pay $1.1 million dollars in compensatory damages and an additional $250,000 in punitive damages.[7]

This article seeks to examine the criminal and tortious liability arising from mosh pits at concerts and the potential defendants named in such an action under

[1] J.D. candidate, Queen's University, Canada. I would like to thank Lynne Hanson of Queen's University Faculty of Law for her guidance and insights and for encouraging my passion for tort law. Thanks to everyone in the pit who has moshed with me and brought joy and excitement into my life.

[2] Jon Campisi, 'Concert-goer injured during Fishbone stage dive awarded $1.4 million' (*The Pennsylvania Record*, 19 February 2014) http://pennrecord.com/news/12870-concert-goer-injured-during-fishbone-stage-dive-awarded-1-4-million> accessed 19 July 2016.

[3] Kyle McGovern 'Fishbone Owe $1.4 Million for Stage-Diving on Fan' (*Spin*, 14 February 2014) <http://www.spin.com/articles/fishbone-stage-dive-lawsuit> accessed 18 July 2016.

[4] ibid.

[5] ibid.

[6] ibid.

[7] ibid.

Canadian law. These defendants can include the owner of a stadium or club where the event takes place, the event coordinator who is occupying the venue, the security, and the patrons themselves. First, this article briefly outlines the evolution and significance of the practice of moshing at concerts. Second, this article will analyse the legality of moshing under American and Canadian law. Third, this article will then identify the potential tortious defendants in an action arising out of a mosh pit incident, how tort law applies to each defendant, and what defences, if any, can be raised. Finally, this article investigates who is liable under Canadian law in a mosh pit incident. The short answer to this question is 'everyone'.

A. What is Moshing?

Moshing is a term used in the punk and metal communities that became synonymous with the frenzied collective form of dancing often seen at concerts. The practice evolved from the 1970's practices of slam dancing that reflected the punk community's message of 'stay away.'[8] The term 'moshing' was not coined to describe the practice until the Washington, D.C., band, Bad Brains, started using 'mash' or 'mash it up' in their lyrics and stage shows. Due to the thick Jamaican accent of vocalist Paul Hudson, the crowd misheard the word 'mash' as 'mosh'.[9] Moshing has been aptly described by one sociologist as 'a huge group fight, except no one's fighting.'[10]

Moshing became an integral part of the concert experience as it allowed bucking social norms, through the release of pent-up frustrations, and the fuelling of a strong communal tradition within a socially acceptable level of violence.[11] Dr. Thomas Hawley, a professor at Eastern Washington University, describes moshing as an outlet for the desire of the will's struggle against what opposes it, in this case dissonance or mental conflict. He says that this state *requires* an outlet such as physical movements of the body. He goes further to say that this struggle against dissonance is not merely a musical phenomenon but rather '…an ontological and phenomenological experience, an explicit and abusive confrontation with all that is terrible in existence.'[12] In short, moshing cannot simply be dismissed as chaotic dancing; for some it is a therapeutic and life affirming exercise.[13]

[8] Joe Ambrose, *Moshpit*: *The Violent World of Mosh Pit Culture* (Omnibus Press 2001) 1.

[9] Gabrielle Riches, 'Embracing the Chaos: Mosh Pits, Extreme Metal Music and Liminality' (2011) 15 For Cultural Research 315.

[10] Craig T. Palmer, 'Mummers and Moshers: Two Rituals of Trust in Changing Social Environments' (2005) 44 Ethnology 147, 154.

[11] Riches (n 9) 316.

[12] Thomas Hawley, 'Dionysus in the Mosh Pit: Nietzschean Reflections on the Role of Music in Recovering the Tragic Disposition' (Paper delivered at the Annual Meeting of the Western Political Science Association, San Francisco, CA, 1-3 April 2010), <https://www.researchgate.net/publication/228277848_Dionysus_ in_the_Mosh_Pit_ Nietzschean_Reflections_on_the _Role_of_Music_in_Recovering_the_Tragic_Disposition>, 33.

[13] ibid 34.

Moshing usually takes place in the semi-circular space in front of the stage but can often extend to the entire arena or standing area where the event is held.[14] A mosh pit can also quickly change to a 'circle pit' as the song or speed of the set changes. A circle pit is comprised of a large number of people running in a circle, sometimes holding onto one another to maintain balance. As the music speeds up so do the participants. Circle pits are generally a good humoured and joyful alternative for when things become too aggressive or heated in the pit.[15]

Depending on the community or 'scene', there are varying levels, or common codes, of conduct shared by participants, known as 'Pit Etiquette'. For instance, consensual jostling and good humoured horseplay is not forbidden but sexual harassment and trampling of fallen members is forbidden.[16] Even with these rules in place, however, mosh pits have grown to as large as 50,000 people at one time and accidents do happen. Minor injuries such as broken noses or sprained ankles are the norm. To treat this, some cities, such as San Francisco, boast a 'Rock Medicine' programme devoted entirely to dealing with mosh-based injuries.[17] Not everyone escapes the pit relatively unscathed, however. In 1994, a 21 year old died from injuries sustained at a Motörhead show in London and 2 participants at a Sepultura and Pantera concert became paraplegics as a result of a mosh pit incident.[18] In June 2000, 9 youths died at the Roskilde Festival in Denmark and in 1999 alone it is estimated that 5,691 concert attendees were injured.[19] Many participants' response to these injuries was a simple message that reinforces the consensual nature of moshing: '[i]f you don't want to get injured, don't go in the pit.'[20]

There is currently no Canadian tort law that responds directly to the practice of moshing. There is, however, a line of cases from the United States, concerning a variety of defendants, that address injuries sustained in mosh pit incidents. A brief examination of American legal response and the limited Canadian criminal law response to moshing gives an outline to the currently somewhat barren legal landscape. This examination finds that moshing is not *prima facie* criminal in the appropriate circumstances. This article then outlines the occupier's statutory duties, under the *Occupiers' Liability Act* and the *Liquor Licence Act*, and the duty of care owed by the occupier/venue/organiser to patrons under the law of negligence. This article analyses the defence of *volenti non fit injuria* to a claim in negligence.

This article then addresses the potential liability of the security at an event where the security is independently contracted. Patrons' liability in battery,

[14] Ambrose (n 8) 2.
[15] ibid 2–3.
[16] ibid 3.
[17] ibid 4.
[18] ibid.
[19] Cecily Lynn Betz, 'The Dangers of Rock Concerts' (2000) 15 Intl J for Pediatric Nurses & Professionals 341.
[20] Ambrose (n 8) 4.

negligence, and negligent battery are then addressed in turn. This article then discusses the defence of contributory negligence. Finally, this article concludes by broadly outlining the potential liability of all parties that arises from injuries sustained in mosh pits at concerts.

2. Moshing: The American Response

While there is very little Canadian case law relating to moshing, the American legal position towards moshing has developed quickly in recent years. To discourage incidents like the aforementioned story of Kimberly Myers, some American cities, such as Boston, have formally banned moshing and slam-dancing, arguing that it is 'dangerous behaviour' that constitutes a 'public safety hazard.'[21] The House of Blues was cited by police when their security did not break up a mosh pit at a Flogging Molly concert.[22] The major music tour known as 'Warped Tour' has done the same by hanging explicit banners that read 'You Mosh, You Crowd Surf, You Get Hurt, We Get Sued, No More Warped Tour'.[23] It is unknown if this has any deterrent effect.

Americans have also taken the unprecedented step to sue not only the location, event organisation, and the security, but to sue the band members themselves for unintentional torts.[24] In *Adams v Metallica*, a plaintiff, Adams, sustained chest trauma inflicted by violent fans in a mosh pit that he voluntarily joined.[25] His claim against the heavy metal band Metallica rested upon a claim of negligent supervision and a failure to warn. He argued that Metallica incited the crowd to mosh and should have anticipated how fans would react to the music of the opening act, 'Suicidal Tendencies'. Based upon this, he tried unsuccessfully to be joined as an intervenor on a similar lawsuit to avoid duplicate discoveries.[26] The main action he sought to join was between a plaintiff named Keith 'Crazy Indian' Philips and Metallica.[27] While within the crowd, Philips volunteered to be launched into the air and then caught by a group of thirty people multiple times. He was intoxicated and acted erratically after drinking from a blue bottle containing unknown contents.[28] Imitating another participant, Philips started spinning while airborne above the crowd. The crowd below him panicked and scattered fearing for their own safety.

[21] Natalie Musumeci 'Boston Police Crackdown on Mosh Pits' (*NBC Bay Area*, 16 March 2012) <http://www.nbcbayarea.com/news/weird/NATL-Boston-Police-Crackdown-On-Mosh-Pits--142945935.html> accessed 23 April 2015.

[22] ibid.

[23] Jason MacNeil 'Warped Tour Tries To Ban Moshing, Crowd Surfing (Which Is Not Very Punk Of Them)' (*Huffington Post*, 21 June 2014) http://www.huffingtonpost.ca/2014/06/21/warped-tour-bans-moshing-crowd-surfing_n_5516336.html accessed 19 July 2016.

[24] *Adams v. Metallica*, 143 Ohio App (3d) 483 (1st App Dist 2001).

[25] ibid 486.

[26] ibid 492.

[27] ibid 485.

[28] ibid.

Philips fell headfirst into the ground, damaging his spine and rendering him a paraplegic.[29] The action was later settled.[30]

Currently there are not Canadian decisions that parallel the legal approach developed in American courts. There is, however, a Canadian legal framework in place that this article applies to provide possible outcomes and obstacles from a Canadian legal perspective. While it does not mirror the American approach, it does provide some parallels in tort liability.

3. Is Moshing Criminal?

There is no Canadian legislation or common law that addresses the legality of mosh pits. The legality of moshing was indirectly addressed in *R. v J.D.C.*, a case concerning the wilful obstruction of an officer in the execution their duty following an altercation in a mosh pit at a concert concerning the defendant.[31] One of the issues in the case was whether the accused's behaviour in the mosh pit amounted to a disturbance under s. 175(1) of the *Criminal Code of Canada*.[32] Section 175(1) of the *Criminal Code* states:

> Every one who *(a)* not being in a dwelling-house, causes a disturbance in or near a public place,
> (i) by fighting, screaming, shouting, swearing, singing or using insulting or obscene language
> (ii) by being drunk, or
> (iii) by impeding or molesting other persons,
> (*b*) openly exposes or exhibits an indecent exhibition in a public place,
> (*c*) loiters in a public place and in any way obstructs persons who are in that place…
> is guilty of an offence punishable on summary conviction.[33]

In *R. v J.D.C.*, the accused entered a mosh pit at a concert and was punched in the face. The accused returned a blow and was placed under arrest for causing a disturbance.[34] In considering whether the accused's actions amounted to

[29] ibid.
[30] ibid.
[31] *R v J.D.C.*, 2009 ABPC 346, [2009] AJ No. 1273 (QL).
[32] ibid [39]–[49].
[33] Criminal Code, RSC 1985, c C-46, s 175(1).
[34] *R v J.D.C.* (n 32) [13].

a disturbance under s. 175(1) of the *Criminal Code*, Judge Redman adopted the following dicta of Allen J. in *R v Edwards*:

> 'The public has a collective right to peace and tranquillity in a public place. This right must be balanced against the right of the individual to express himself or herself. Some disruption of the peace and tranquillity of a public place must be tolerated. A determination whether the public right to peace and tranquillity has been disturbed is a factual determination to be made by the trier of fact recognizing the competing interests. The disturbance is of the public's use of a public place and not the disturbance of an individual's mind. The intensity of the activity and its effect on the degree and nature of the peace that is expected to prevail at the particular time must be considered. The trier of fact must find that there is an externally manifested disturbance of the public peace in the sense of interference with the ordinary and customary use of a public place. The disturbance may consist of the impugned act itself or a consequence of the impugned act.'[35]

Judge Redman concluded that '[t]he mere act of moshing aggressively does not seem to me to be causing a disturbance in the context of a mosh pit at a rock concert where the music is loud, the bodies are close and people are flinging themselves around at each other.'[36] Judge Redman found that the officer had no reasonable and probable grounds to believe the accused was causing a disturbance within the meaning of s. 175(1) and the accused was acquitted of all charges.[37]

While mosh pits have not been the subject of much criminal litigation in Canada, injuries sustained in a mosh pit may be the result of other offences under the *Criminal Code*, such as assault.[38] Section 265(1) of the *Criminal Code* states that a person commits an assault when:

a. without the consent of another person, he applies force intentionally to that other person directly or indirectly
b. he attempts or threatens, by an act or a gesture, to apply force to another person, if he has, or causes that person to believe on reasonable grounds that he has, present ability to effect his purpose; or
c. while openly wearing or carrying a weapon or an imitation thereof, he accosts or impedes another person or begs.

What might otherwise be considered an assault will not be considered a criminal assault for the purposes of s. 265(1) where there is a 'social utility', as

[35] *R v Edwards*, 2004 ABPC 14 [89].
[36] *R v J.D.C.* (n 32) [49].
[37] ibid [49], [84].
[38] s. 265, Criminal Code.

discussed in *R. v Jobidon*.[39] In *Jobidon*, the Court recognised that exceptions were created for assaults that have a 'social utility' but failed to define what 'social utility' means. This 'social utility' test was echoed in *R v Adamiec*, where it was held that a sport – in this case, ice hockey – has a 'social utility in providing exercise and entertainment' and plays an important of Canadian identity and culture.[40] Following this admittedly uncertain criteria for the 'social utility' test and the dicta of Allen J. in *Edwards*, it is fair to say that if moshing were not in an appropriate location or if the intensity of moshing was too extreme, moshing and injuries sustained as a result of moshing may result in criminal liability.

4. Moshing and Occupier's Liability

In Canada, the occupier's duties are addressed both by legislation as well as by the common law. This Section will specifically address liability arising under legislation in Ontario, namely the *Occupiers' Liability Act* and the *Liquor Licence Act*, as well as *Regulation 719 Licences to Sell Liquor*.

A. Occupiers' Liability Act

The *Occupiers' Liability Act* outlines the occupier's duty in section 3(1):

> An occupier of premises owes a duty to take such care as in all the circumstances of the case is reasonable to see that persons entering on the premises, and the property brought on the premises by those persons are reasonably safe while on the premises.[41]

The duty of care applies to both the premises themselves and any activities carried out on the premises.[42] Section 4(1) of the *Occupiers' Liability Act* narrows the duty of care to exclude 'risks willingly assumed':

> The duty of care provided for in subsection 3(1) does not apply in respect of risks willingly assumed by the person who enters on the premises, but in that case the occupier owes a duty to the person to not create a danger with the deliberate intent of doing harm or damage to the person or his property and to not act with reckless disregard of the presence of the person or his property.[43]

[39] *R v Jobidon* [1991] 2 SCR 741.
[40] *R v Adamiec* 2013 MBQB 246, [24]–[25].
[41] Occupiers' Liability Act, RSO 1990, c O2, s 3(1) [Occupiers' Liability Act].
[42] ibid s 3(2).
[43] Occupiers' Liability Act 1990, s 4(1).

The duty of care owed by an occupier under section 3(1) and section 4(1) of the Ontario *Occupiers' Liability Act* was discussed in *Waldick v Malcolm*.[44] Blair J.A. noted that the duty under section 3(1) is not absolute and that occupiers are not insurers liable for any damages suffered by persons entering their premises.[45] Blair J.A. further noted that trier of fact must determine what standard of care is reasonable and whether it has been met.[46] When discussing the duty of care arising under section 4(1), Blair J.A. found that section 4(1) required not only knowledge of the risk but also physical and legal acceptance of the risk by the visitor or patron.[47] In other words, this supports a codification of the *volenti* doctrine in Canada,[48] discussed in greater detail in Section 6 of this Article. Unless mosh pit participants are proven to be knowledgeable and accepting of the risks of entering the mosh pit on the occupier's premises, there may therefore be a duty owed on the part of the arena or stadium owner or occupier.

Occupiers' liability can potentially also extend to anyone who rents out the premises. In *Jacobson v Kinsmen Club of Nanaimo*, the defendant society rented out a curling club to hold a beer garden.[49] The roof of the club was supported by a series of I-beams which were accessible from the ground. The plaintiffs entered, consumed alcohol, and then began climbing the I-beams to the amusement of the other patrons. One patron lost his grip on the I-beam and fell thirty feet onto an unsuspecting patron, injuring him.[50] The defendant society was found to be a liable occupier under the British Columbian *Occupiers' Liability Act*.[51] Under the British Columbian *Occupiers' Liability Act*, liability extends to an event organiser or coordinator who rents out the stadium or arena and controls the premises for the purposes of a concert.[52] Failure to meet this duty of care can lead to a finding of negligence.

B. Liquor Licence Act

Where a venue which hosts musical acts serves alcohol, the *Liquor Licence Act* imposes additional duties for the occupier towards persons on the premises.[53] Section 39 of the Act outlines the civil liability of the occupier as an alcohol vendor and extends occupiers' liability to all the occupier's agents or employees if their sale of

[44] *Waldick v Malcolm* [1989] OJ No. 1970.
[45] ibid [18].
[46] ibid.
[47] ibid [32]–[40].
[48] ibid [32].
[49] *Jacobson v Kinsmen Club of Nanaimo*, (1977) 71 DLR (3d) 227 (QL).
[50] ibid [9].
[51] Occupiers Liability Act (British Columbia) 1990, s 1(b).
[52] ibid.
[53] Liquor Licence Act, RSO 1990, c. L. 19 [Liquor Licence Act].

alcohol results in a level of intoxication that makes a person a danger to others or themselves.[54]

Under section 39, the licence holder can be liable for the injuries caused by drunk mosh pit participants to each other. This section cannot be invoked, however, if the plaintiff is blameworthy, thus limiting its application.[55] The Court in *Sambell v Hudago Enterprises* added that '[this] duty on tavern owners is not absolute or unbounded but they must act reasonably to protect against the risk apprehended. What is reasonable depends on the circumstances and the magnitude of the risk.'[56] This position was later complicated by *Hague v Billings*, which states that if the requirements of section 39 of the Act are met, absolute liability is imposed and the issue of causation becomes irrelevant.[57] Somers J subsequently addressed *Hague v Billings* in a motion for summary dismissal.[58] He muddied the waters by stating that 'the principles respecting liability based on s. 39 of the [Act] are not entirely settled.'[59] The court went on to infer that the tort requirement of causality does apply in the traditional manner.[60] Due to this disagreement on the bench, it is somewhat unclear which analysis section 39 requires.

C. Regulation 719 Licences to Sell Liquor

The licence holder— usually the occupier—is also bound by *Regulation 719 Licences to Sell Liquor*. Section 45 of this regulation states:

> The licence holder shall not permit drunkenness, unlawful gambling or riotous, quarrelsome, violent or disorderly conduct to occur on the premises or in the adjacent washrooms, liquor and food preparation areas and storage areas under the exclusive control of the licence holder.[61]

This section appears to impose an obligation upon the occupier to deter moshing in a place where alcohol is sold. Alternatively, the licence holder would bear the onus to prove that moshing is not violent or disorderly conduct. The common law has injected a level of reasonableness into this section. Section 45 must be interpreted 'reasonably in accordance with its plain language and the practicalities of the context in which it is applied.'[62] With this added gloss of reasonableness,

[54] Liquor Licence Act, s. 39.
[55] *Menow v Honsberger* [1974] SCR 239, [11]–[12].
[56] *Sambell v Hudago Enterprises Ltd* [1990] OJ No. 2494, [45].
[57] *Hague v Billings* [1993] OJ No. 945, [15].
[58] *Haughton v Burden* [2001] OJ No. 4704 [24].
[59] ibid.
[60] ibid [25]–[26].
[61] Liquor Licence Act RRO 1990, Reg 719, s 45 (Reg 719).
[62] *Horseshoe Valley Resort Ltd v Ontario (Alcohol & Gaming Commission)* [2005] OJ No. 5895, [14].

the occupier's duty owed to the moshing patron is ambiguous at best. Nevertheless, it important to recognise that, ambiguous though it may be, there is a duty of care owed by an occupier to a moshing patron that, if breached, can lead to a finding of negligence.

5. Moshing: An Action in Negligence?

Would it be possible to for an injured party sue the venue, security, or patrons for failing to prevent injury in a mosh pit? In answering this question, this Section will next focus on the duty of care, the standard of breach, and causation.[63] This Section will next turn to the liability in negligence of the venue specifically.

A. Duty of Care

In order to be found liable, the defendant must firstly owe a duty of care to the patrons. The test for the existence of a duty of care in the tort of negligence is the two-stage *Anns* test, as endorsed by the Supreme Court of Canada in *Cooper v Hobart*.[64] At the first branch of the test, two questions arise:

> (1) was the harm that occurred the reasonably foreseeable consequence of the defendant's act? and (2) are there reasons, notwithstanding the proximity between the parties established in the first part of this test, that tort liability should not be recognised here?[65]

The proximity analysis focuses on factors arising from the relationship between the plaintiff and the defendant[66]—in this case, between the injured patron and the venue owner or occupier—looking to their interests while participating within the mosh pit, including expectations, representations, or reliance.[67] For example, depending on the patrons' state of mind and knowledge of mosh pits and the venue, patrons may or may not have expectations, reasonable or otherwise, that they will not be injured or jostled. Once foreseeability and proximity are established at the first stage, a *prima facie* duty of care arises.[68] The second stage of the *Anns* test is concerned with whether there are any residual policy considerations outside the relationship of the parties which ought to negate or limit the scope of the duty, the class of persons to whom the duty is *prima facie* owed, or indeterminate damages

[63] This article does not examine the requirement 'proximate cause' since it will almost always be met in the setting of a mosh pit.
[64] *Cooper v Hobart*, 2001 SCC 79, [2001] 3 S.C.R. 537.
[65] ibid [30].
[66] ibid.
[67] ibid [33].
[68] ibid.

which may result.[69] The second stage is generally only applied if the situation is novel.[70]

B. Standard of Care

The second requirement for a finding of negligence is the breach of the standard of care. All parties are held to a *standard* of care if a duty of care is established. In assessing the standard of care, the starting point is a legal fiction known as the 'reasonable man'. The reasonable man is a 'mythical creature' who sets the appropriate level of conduct of all other persons.[71] He possesses no superhuman traits, skills or intelligence, rather he is a 'reasonable and prudent man.'[72] Once the correct standard is ascertained, the defendant's actions are compared to it. If the defendant fails to meet the standard of care by their actions, it will constitute a breach of the standard.[73] For example, if the security team ignored an injured patron in a mosh pit and failed to remove them or provide them with medical attention because they were having a beer or watching the band, the security team would have breached the standard of care expected of a reasonable security worker. Similarly, if the occupier failed to ensure the area was clean of broken glass or other debris that could harm a patron, the occupier would also breach the standard of care.

C. Causation

The third requirement is causation. The tort of negligence can be caused by the actions of one or of a group of tortfeasors. The test for establishing causation is the 'but for' test.[74] The test places the burden upon the plaintiff to demonstrate that, on a balance of probabilities, but for the negligent act or omission of the defendant(s), the plaintiff would not have been injured.[75] The 'but for' test requires a 'substantial connection between the injury and the defendant's conduct' to shield innocent defendants from unconnected causes.[76]

If an injury were sustained in a mosh pit and the injured person was seeking to sue a particular party, the causation requirement would be established if *but for* the actions of the defendant, the injury would not have occurred. If the security had acted swiftly or effectively would there be an injury? What if the occupier had cleared the area of spilled beer or ice? If the purported act or omission is integral

[69] ibid [37]–[38].
[70] ibid [39].
[71] *Arland and Arland v Taylor*, [1955] OR 131 [29].
[72] ibid.
[73] *Clements v Clements*, [2012] 2 SCR 181, 2012 SCC 32, [6].
[74] *Resurfice Corp v Hanke*, [2007] 1 SCR 333, 2007 SCC 7, [18]–[21] [*Hanke*].
[75] *Blackwater v Plint*, [2005] 3 SCR, 2005 SCC 58 [78].
[76] *Hanke* (n 75) [23].

to the chain of causation and its removal might, on the balance of probabilities, have prevented the damage, then the causation requirement is met. If there is a relationship of proximity sufficient to attract liability and the harm was reasonably foreseable, then the defendant—be it the venue owner or occupier, security, or patrons—can be found liable under the law of negligence for the injury sustained.

D. The Venue's Liability in Negligence

In addition to any statutory obligations,[77] there is a common law duty created by the 'special relationship' between patrons and venues/organisers, the breach of which may result in the venue being found liable in negligence. This common law duty has evolved through a series of cases which will be discussed in this Section.

In *Hessie v Laurie*, the plaintiff, a patron, was assaulted by a second patron after coming to the aid of an employee who was attempting to remove a violent and intoxicated patron.[78] The plaintiff sought to hold both the intoxicated patron and the tavern liable for his injuries. He argued that the establishment owed him and the other patrons reasonable care in protecting them from other patrons. Riley J described the elevated standard of care for patrons who are served alcohol as 'anxious care', a standard that is subjective to the locale, the type and character of its usual patrons, the size of its operations, and what occurrences ought reasonably to be anticipated and guarded against.[79]

In *Crocker v Sundance Northwest Resorts Ltd.*, the Supreme Court of Canada stated that '[t]he common thread running through [the case law] is that one is under a duty not to place another person in a position where it is foreseeable that the person could suffer injury.'[80] In *Crocker*, a ski resort was found liable for the neck injury rendering the plaintiff a qaudriplegic because the resort allowed a patron to participate in a tubing competition after serving him to the point of intoxication.[81] The Supreme Court noted that it was relevant to 'relate the probability and gravity of injury to the burden that would be imposed upon the prospective defendant in taking measures.'[82] The Court found the nexus between Sundance Resort and Crocker too close for Sundance to be a 'stranger to Crocker's misfortune.'[83] Sundance had a responsibility to prevent intoxicated persons from participating in a dangerous sport.[84] The Supreme Court employed the same reasoning in *Stewart v Pettie* and clarified that, even with the existence of a 'special relationship',

[77] See section 4A–4C.
[78] *Hessie v Laurie* (1962), 35 DLR (2d) 413 [22].
[79] ibid [26].
[80] *Crocker v Sundance Northwest Resorts Ltd*, [1988] 1 SCR 1186, [21].
[81] ibid [39].
[82] ibid [20].
[83] ibid [23].
[84] ibid [24].

there is no positive duty on the operator unless there is a foreseeable risk.[85] An occupier's failure to meet this duty of care can lead to a finding of negligence, but a relationship between parties alone is not sufficient.

The Supreme Court restated this duty in 2006 in *Childs v. Desormeaux*.[86] The Court found that the commercial relationship between patron and tavern creates a duty of care for three reasons. Firstly, commercial hosts are trained and expected to monitor alcohol consumption.[87] Secondly, the sale and consumption of alcohol is strictly regulated by legislatures.[88] Third, there is an incentive to overserve in hopes of maximising profit: 'the benefits of over-consumption go to the tavern keeper alone, who enjoys large profit margins from customers whose judgment becomes more impaired the more they consume. This perverse incentive supports the imposition of a duty to monitor alcohol consumption in the interests of the general public.'[89] The duty of a venue to protect a patrons is thus established where there is foreseeability of harm *and* a 'nexus' or 'special relationship'—usually a commercial relationship—between the venue and patron.[90]

On this analysis, a venue serving alcohol at a concert that is known to have mosh pits is not only likely to have the requisite 'nexus' but is also likely to owe a duty of care to its patrons. If the venue serving alcohol and hosting the event were the same entity, the scenario would closely mirror that of *Crocker*. If it is foreseeable that there will be a mosh pit at the concert, the venue operator may have to guard against a mosh pit; the probability and gravity of injury are both very real and may elevate the standard of care owed by the venue to the patron.

How far does the duty to protect against injury extend? A patron who is reasonably served may still be injured in a mosh pit due to slightly diminished response times. The plaintiff in *Crocker* was clearly drunk, but what happens where a patron is served only one or two drinks? If the patron is able to make informed decisions, will the duty to protect still be established if a mosh pit suddenly becomes rambunctious? Will an injury sustained in a mosh pit always be a foreseeable harm due to a mosh pit's inherently dangerous nature? While *Crocker* and *Childs* propose a useful framework where a patron is blatantly overserved and then courts danger, they fail to provide guidance for the grey areas between the extremes.

6. Volenti

Volenti is a defence to negligence based on the maxim '*volenti non fit injuria*' which means that a person cannot complain of harm consented to within his knowledge and free will. The defence of *volenti* applies where parties give express or implied

[85] *Stewart v Pettie*, [1995] 1 SCR 131, [48]–[50].
[86] *Childs v Desormeaux*, [2006] 1 SCR 643, 2006 SCC 18.
[87] ibid [18].
[88] ibid [19].
[89] ibid [22].
[90] ibid [31]–[34].

consent to assume risks without compensation. In *Dubé v Labar*, Estey J set the bar relatively high for invoking the defence of *volenti*, stating:

> [*V*]*olenti* will arise only where the circumstances are such that it is clear that the plaintiff, knowing of the virtually certain risk of harm, in essence bargained away his right to sue for injuries incurred as a result of any negligence on the defendant's part. The acceptance of risk may be express or may arise by necessary implication from the conduct of the parties, but it will arise, in cases such as the present, only where there can truly be said to be an understanding on the part of both parties that the defendant assumed no responsibility to take due care for the safety of the plaintiff, and that the plaintiff did not expect him to.[91]

Under the doctrine of *volenti*, it may be possible for a venue or an organiser to release itself from liability where patrons sign a binding document prior to the concert or where the patrons agree to terms and conditions upon purchasing their tickets. In *Dimopoulos v Thiessen*, the defendant negligently crosschecked the plaintiff in the mouth during a ball hockey game.[92] The court found that the plaintiff assumed the risk when he registered his team and signed the 'sign up sheet' which contained a release and waiver.[93]

This precept that consensual injuries are not actionable often applies to blows given or injuries received in fair play and not maliciously. In *Agar v Canning*, the Court drew a sharp division between a blow struck in the course of sport, which it found to be acceptable, and a blow struck in anger or maliciously, which it found to be a battery.[94] The court held that what a player does in the heat of sport should not be 'judged by standards suited to polite social intercourse.'[95] This defence extends, however, beyond organised sport. In *Wright v McLean*, four young boys threws balls of mud or clay at each other for sport.[96] One of the boys accidentally threw a stone mistaking it for mud and injured the plaintiff. The presiding judge found no civil liability, relying on consent to justify his finding.

On this analysis, could moshing be included under the wide umbrella of sport? Moshing is at the very least dancing, which is often described as a sport or at least grouped with sport.[97] It provides the participants with exercise and entertainment and holds cultural value for many subcultures. The venue, the occupier, or the organiser can similarly invoke the defence of *volenti* in the same fashion of a release

[91] *Dubé v Labar*, [1986] 1 SCR 649 [6].
[92] *Dimopoulos v Thiessen Signing Doc*, 2009 BCPC 140 [3].
[93] ibid [16].
[94] *Agar v Canning*, [1965] 54 WWR 302, [3]–[4].
[95] ibid [7].
[96] *Wright v McLean* [1956] WWR 305 [1].
[97] *Bonenfant v Campagna*, [1977]16 NBR (2d) 544, [1].

and waiver. There are, however, exceptions whereby a venue or an organiser will not be able to invoke the defence of *volenti*, even where a release and waiver has been signed.

These exceptions were outlined by McLachlin J in *Karroll v Silver Star Mountain Resorts Ltd.*[98] The defendant ski resort was sued for negligently failing to ensure that the course was empty of other skiers when the plaintiff descended.[99] The plaintiff relied on the defendant's assurance and subsequently collided with another skier.[100] The resort relied upon a release and indemnity which the plaintiff signed prior to participating.[101] The court outlined three circumstances in which a defendant cannot rely upon a written agreement of this kind: first, where the contract is signed by mistake;[102] second, where the signing is induced by fraud or misrepresentation,[103] and; third, where the provider of the contract is aware that it is misunderstood or mistaken by the signor.[104]

An occupier can, in theory, waive their occupier's liability but the occupier has to do so carefully in a way that all participants understand and acknowledge that the occupier is waiving its liability for any injuries sustained. For example, a waiver might be digitally signed by a patron upon purchase of tickets to an event. This waiver might then serve as evidence for raising the defence of *volenti* to a claim in negligence against the venue or the organiser. The venue or organiser will still be vulnerable however, to the three exceptions enunciated in *Karroll v Silver Star.*

7. Security's Liability

Many venues employ some level of security, such as 'bouncers', to control the premises and to protect the patrons from acts of aggression or danger. If the security are employees of the occupier, they are generally shielded by vicarious liability; it is the occupier, and not the employees, who will be held vicariously liable. The security workers can be held liable, however, where they are independently contracted for the event.

[98] *Karroll v Silver Star Mountain Resorts Ltd* [1988] 33 BCLR (2d) 160, 1988 CanLII 3094 (BS SC).
[99] ibid.
[100] ibid.
[101] ibid.
[102] ibid.
[103] ibid.
[104] ibid.

Where security workers are independently contracted, their liability is severed from the occupier, assuming that the occupier acted reasonably. Security workers will owe a duty to patrons, the breach of which can lead to a claim of negligence by the patrons that they were employed to protect:

> Where damage to any person or his or her property is caused by the negligence of an independent contractor employed by the occupier, the occupier is not on that account liable if in all the circumstances the occupier had acted reasonably in entrusting the work to the independent contractor, if the occupier had taken such steps, if any, as the occupier reasonably ought in order to be satisfied that the contractor was competent and that the work had been properly done, and if it was reasonable that the work performed by the independent contractor should have been undertaken.[105]

The duty and standard of care owed by security to those they are hired to protect was addressed by Cromwell J in *Fullowka v Pinkertons of Canada Ltd*.[106] In *Fullowka*, during a mine strike, a disgruntled striker evaded security and set a bomb that resulted in the death of nine miners.[107] Cromwell J outlined the test for foreseeability as 'whether the harm would be viewed by a reasonable person as being very likely to occur'.[108] He then examined the proximity between the security and the miners to see if there had a positive duty to act. Cromwell J, citing *Childs*,[109] noted that there were at least three factors which may identify a positive duty to act:

> The first is that the defendant is materially implicated in the creation of the risk or has control of the risk to which others have been invited. The second is the concern for the autonomy of the persons affected by the positive action proposed. As the Chief Justice put it: 'The law ... accepts that competent people have the right to engage in risky activities ... [and] permits third parties witnessing risk to decide not to become rescuers or otherwise intervene.' The third is whether the plaintiff reasonably relied on the defendant to avoid and minimize risk and whether the defendant, in turn, would reasonably expect such reliance.[110]

[105] Occupiers' Liability Act (n 42) s 6(1).
[106] *Fullowka v Pinkerton's of Canada Ltd*, [2010] 1 SCR 132, 2010 SCC 5.
[107] ibid [4]–[9].
[108] ibid 21.
[109] *Childs* (n 87) [31]–[46].
[110] *Fullowka* (n 107) [27].

This analysis can be applied to a security team at a concert with mosh pits. In establishing whether a *prima facie* duty of care exists between the security and patrons, there is sufficient foreseeability as harm could be 'viewed by a reasonable person as being very likely to occur.'[111] Due to the inherently dangerous nature of mosh pits, this is easily met. Additionally, proximity may also be met following an application of the three factors. First, the security are not materially implicated in the creation of the risk but their sole purpose is control of the venue and the risk contained within. Second, the patrons are involved in risky activities, which are not excluded nor do they require third parties to intervene. Third, the patrons could reasonably rely on security for their safety as this is the purpose of their employment. On this analysis, it is possible that the security will owe a duty of care towards the patrons of an event but the existence of the duty will most likely depend on the time and place of the event. In *Fullowka*, Pinkertons did not breach their standard of care as they were understaffed (against their own urges) and unable to meet the appropriate standard of care through no fault of their own.[112] In a mosh pit scenario, a properly staffed security team which fail to discharge the proper standard of care to patrons may be held liable for injuries sustained by patrons.

8. Patrons' Liability

A patron injured at a concert could seek to sue other patrons for the injuries sustained. The patrons who enter the mosh pit hold the lion's share of responsibility for their conduct. Patrons can be held liable in negligence, but it is more likely that they will be sued in a tort that has an element of intention. There are three torts that can lead to patron liability: battery, negligence, and negligent battery.

A. Battery

While American case law uses battery and assault separately, *Gambriell v Caparelli*, established that the distinction between assault and battery have been blurred in criminal matters and eliminated in civil matters.[113] For this reason, there is no need to address these two torts separately.

Bettel v Yim defined battery as 'the intentional infliction upon the body of another of a harmful or offensive contact.'[114] In *Bettel v Yim*, the plaintiff started a small fire in the defendant's store. The defendant tried to coerce a confession from the plaintiff and shook him two or three times.[115] During this shaking, the plaintiff's

[111] ibid [21].
[112] ibid [80].
[113] *Gambriell v Caparelli* (1974) 7 OR (2d) 205, 54 DLR (3d) 661 [13].
[114] *Bettel Et Al v Yim*, [1978] DLR (3d) 543.
[115] ibid [6].

nose accidentally struck the defendant's head causing his nose to bleed.[116] The defendant was found liable for the plaintiff's injuries despite the fact the injuries were not intended; the damages were a result of the defendant's intentional touching, resulting in the defendant's responsibility.[117]

A battery-free mosh pit is somewhat of an oxymoron due to the slam dancing that occurs in a mosh pit, similar to checking in an ice hockey game. If any patron in a mosh pit intentionally touches another patron and the contact resuls in harm, it may amount to battery. Even if a patron simply intends to bump and jostle with others harmlessly, but misjudges their strength and injures another patron, they could be liable for battery.

B. Negligent Battery

Mosh pits are generally not a place of calculated movement so there is a strong possibility of negligent battery. In both *J.A.S. v Gross*[118] and *Non-Marine Underwriters, Lloyd's London v Scalera*,[119] the Supreme Court of Canada deferred to Lewis Klar's definition of negligent battery:

> A negligent battery exists when the defendant causes a direct, offensive, physical contact with the plaintiff as a result of negligent conduct. The defendant's negligence consists of unreasonably disregarding a foreseeable risk of contact, even though the contact was neither desired nor substantially certain to occur.[120]

Negligent battery is not often pleaded but it does still survive as a cause of action, as seen in *Kinkade*, a case in which a patron at a club was shot in the leg by a club employee following an altercation.[121] The claim of negligent battery was not directly addressed since it was subsumed into the claim for negligence.[122] Negligent battery could, however, still be pleaded by an injured patron for injuries suffered in a mosh pit depending on the circumstances.

In the chaos of a mosh pit, a negligent battery is entirely possible, if not more likely than anywhere else. The Court in *Gross* states that negligent battery requires harm as a result of disregarding a 'foreseeable risk of physical contact'.[123] All that is required in a mosh pit is for a patron to throw their body and limbs in close contact

[116] ibid [7].
[117] ibid [37].
[118] *J.A.S. v Gross*, 2002 ABCA 36.
[119] *Non-Marine Underwriters, Llyod's of London v Scalera* [2000] 1 SCR 551, 200 SCC 24.
[120] Lewis Klar, *Tort Law*, (2nd edn Carswell, 1996) 47.
[121] *Kinkade v 947014 Ontario Inc c.o.b. as The Silver Dollar*, 2014 ONSC 1599.
[122] ibid [47]–[48].
[123] *J.A.S. v Gross* (n 119).

knowing but disregarding the risk this contact poses to other patrons. Similar to battery, negligent battery is arguably unavoidable in a mosh pit since it goes to the very nature of the activity.

C. Contributory Negligence

Contributory negligence is a defence to a claim in negligence that depends on the negligence of a party, usually the plaintiff, authoring their own injury.[124] A plaintiff always owes a duty to care of themselves and all that is necessary to raise it as a defence is 'that the injured party did not take reasonable care of himself and contributed, by this want of care, to his own injuries.'[125] For example, in *Glanville v Moberg*, a plaintiff became intoxicated and failed to wear a seatbelt when he was involved in a motor vehicle accident.[126] The plaintiff's failure to wear a seatbealt was found to have contributed to a 'substantial portion of the fault' and his liability was assessed at thirty percent.[127]

Where the plaintiff is found to have contributed to the injury or damage caused, the question of apportionment of damages inevitably arises. Section 3 of the *Negligence Act* states that where 'negligence is found on the part of the plaintiff that contributed to the damages, the Court shall apportion the damages in proportion to the degree of fault or negligence found against the parties respectively.'[128] The apportionment of damages is therefore made on the basis of the degree of *fault* found against the parties and *not* causation.[129] The measurement is a question of how far each person deviated from the standard of care, not how much damage they caused.[130] Where it is impractical to determine the measurement of deviation from the standard of care, the parties are found to be equally at fault.[131]

In the case of injuries sustained in a mosh pit, contributory negligence could arise in a variety of permutations between the venue, multiple patrons, and security. For example, patron A is overserved by the establishment at a concert and throws himself recklessly into a mosh pit. Patron A falls causing Patron B to trip, resulting in both Patron A and Patron B being trampled upon in the mosh pit. This incident goes unnoticed by Security Guard C, who is tired and 'resting his eyes' and ought to have prevented the incident in the mosh pit by keeping drunk patrons, such as Patron A, away from the mosh pit.

The liability of each party in the above example will depend on the extent to which each party deviates from their individual standard of care. Security Guard

[124] *Fraser v Ortman*, [1980] AJ No. 629 [9].
[125] ibid.
[126] *Glanville v Moberg*, 2014 BCSC 1336, [12].
[127] ibid [122].
[128] Negligence Act, RSO 1990, c N1.
[129] *Cempel v Harrison Hot Springs Hotel Ltd.*, 1997 CanLII 2374 (BC CA), [1998] 6 WWR 233 [19].
[130] ibid.
[131] Negligence Act, s 4.

C and the venue can both be held partially liable for the injuries of Patron A and Patron B since both deviated from the required standard of care as established in *Crocker* and *Fullowka* respectively.[132] Patron A might be found negligent and to have contributed to his own injuries or damage for drinking too much, recklessly throwing himself into the mosh pit, and failing to assess the situation before acting. If Patron A is found negligent, he will be liable for a portion of his injuries as well as those of Patron B.

9. Conclusion

Moshing is not *prima facie* criminal. In the appropriate setting and with the appropriate intensity of force and supervision, moshing may even be protected by the right to freedom of expression if it takes in an appropriate public place.[133] Moshing is not, however, without its dangers and may, where damage occurs, result in liability of various parties for the damage caused for failure to uphold requisite obligations or standard of care.

The venue and/or event organiser bear(s) a variety of different obligations to patrons. The venue has a statutory duty to keep patrons 'reasonably safe' while they are on the premises pursuant to the *Occupiers' Liability Act*. Where alcohol is sold at the event, the *Liqour Licence Act* saddles the occupiers with an additional level of responsibility.[134] *Regulation 719 Licences to Sell Liquor* ensures that the venue discourages patrons' 'violent or disorderly conduct' but, since it requires a contextual application, it falls short of providing a concrete example. Where the security at the event is independently contracted, there may be a duty owed by the security to the patrons depending on the control of the risk, balanced with autonomy of the person, and reliance upon their intervention.[135] It is possible that patrons may also be liable for injuries sustained by other patrons under the law of battery, negligent battery, or ngliegence, depending on their behaviour in the mosh pit. If the plaintiff consents to the harm, the defendant may be absolved of liability if the activity provides a 'social utility', though it is yet to be determined whether moshing falls under the 'social utility' exception. Similarly, contributory negligence could divide or diminish the liability of all the parties involved depending on the scenario.

The simple answer to the question posed by this article, who can be held liable for injuries sustained in a mosh pit at concert, is 'everyone'.

[132] *Crocker* (n 84); *Fullowka* (n 107).
[133] *Edwards* (n 36) 89.
[134] *Haughton* (n 59) [24]–[26].
[135] *Fullowka* (n 107).

The Conceptual Relationship Between Privacy and Data Protection

Aidan Forde[1]

1. Introduction

RECENT DECISIONS FROM the Court of Justice of the European Union ('CJEU') in *Google Spain*[2], *Digital Rights Ireland*[3] and *Schrems*[4] illustrate an increasingly emboldened stance towards informational privacy and data protection issues by the CJEU.[5] In order for the judicial adjudication of privacy before the CJEU to be effective, it is necessary to pinpoint the inherent value of privacy, its conceptual foundations, and any competing considerations. After the Lisbon Treaty, informational privacy is now recognised as having attained the status of a constitutional right across the EU landscape, finding its constitutional

[1] BCL (Int.), LL.M. (Cantab), Aidan.forde@cantab.net.

[2] C-131/12 *Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González* [2014] All ER (EC) 717.

[3] Joined Cases C-293 & 594/12 *Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources, Minister for Justice, Equality and Law Reform, The Commissioner of the Garda Síochána, Ireland and the Attorney General, and Kärntner Landesregierung, Michael Seitlinger, Christof Tschohl and Others* [2014] All ER (D) 66 (Apr).

[4] C-362/14 *Maximillian Schrems v Data Protection Commissioner* (6 October 2015).

[5] Informational privacy is founded upon personal autonomy and involves protecting and controlling information relating to the individual. It surrounds 'The freedom of the individual to decide on himself is at stake when the individual is uncertain about what is known about him, particularly where what society might view as deviant behaviour is at stake (the chilling effect). The individual therefore has the right to know and to decide on the information being processed about him. At the same time, as a social being, the individual cannot avoid becoming the object of information processing. However, limitations to his basic right are to be accepted only when there is an overriding general interest and where that interest is molded into a law that follows the basic requirements of clarity and proportionality. To protect these principles, a number of safeguards are required: the safeguards consist of data protection principles (correctness, timeliness, purpose limitation, fairly and lawfully obtained), derived rights (access, correction), and organisational safeguards (independent institutions).' BVerfGE 65, 1 ff (1983). Lynskey notes that ' ' Orla Lynskey, 'Deconstructing Data Protection: The "Added-Value" of a Right to Data Protection in the EU Legal Order' (2015) 63 International & Comparative Law Quarterly 569, 590.

basis in article 8 of the Charter of the European Union ('EU Charter'). Under the framework of the European Convention of Human Rights ('ECHR'), informational privacy has been considered under Article 8 of the Convention.[6] The trouble arises, however, in identifying the conceptual foundations of the right to informational privacy under both the EU Charter and the ECHR. Though located in separate provisions of distinct instruments[7], there is, in recent times, an increasing convergence in the conceptual bases upon which the CJEU and European Court Human Rights ('ECtHR') have upheld informational privacy claims.

This article seeks to examine the relationship between the right to informational privacy and the right to data protection under both the EU Charter and the ECHR, utilising the perspectives offered by Paul de Hert and Serge Gutwirth, Lee Bygrave, and Orla Lynskey to critically analyse these two significant rights. Section 1 of this Article will outline the benefit of data protection. Section 2 will outline the perspectives offered by Paul de Hert and Serge Gutwirth, Lee Bygrave, and Orla Lynskey in turn. Section 3 will tie each of these theories together. I will conclude that, in light of the analysis of these theories, that data protection is located on the fringes of privacy and that many of the justifications for data protection overlap with the justifications for privacy.

2. The Benefit of Data Protection

It is an onerous task to identify a unified conceptual understanding of privacy. Controversy surrounds its relationship with data protection.[8] Notwithstanding the battle to identify such a unified conceptual understanding of privacy,[9] elucidating the relationship between privacy and data protection is something of clear benefit to democracy and society. Preventing disproportionate and unlawful intrusion into privacy serves as a shield to totalitarian states.[10] Totalitarianism flourishes when privacy rights are diminished. By contrast, a healthy democratic state will

[6] See 'Internet: Case-law of the European Court of Human Rights' (June 2015), 8 <http://www.echr.coe.int/Documents/Research_report_internet_ENG.pdf> Accessed 23 August 2016.

[7] TJ McIntyre, 'Implementing Information Privacy Rights in Ireland' in Suzanne Egan (ed) *International Human Rights: Perspectives from Ireland* (Dublin: Bloomsbury, 2015) 294.

[8] Juliane Kokott and Christoph Sobotta, 'The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR' (2013) 3(4) International Data Privacy Law 222–228; Raphael Gellert and Serge Gutwirth, 'The legal construction of privacy and data protection' (2013) 29(5) Computer Law & Security Review; Peter Blume, 'Data Protection and Privacy—Basic Concepts in a Changing World' (2010) 56 Scandanavian Studies in Law 297–318.

[9] ibid.

[10] Daniel J. Solove, 'Privacy and Power: Computer Databases and Metaphors for Information Privacy' (2001) 53 Stanford Law Review 1393; Neil M. Richards, 'The Dangers of Surveillance' 2013 126 Harvard Law Review 1934; Hina Sarfaraz, 'Surveillance, privacy and cyber law' (2014) 20 7 Computer and Telecommunications Law Review 189.

enable its citizens to live independent and informed lives.[11] When unlawful and disproportionate interference with the private zone occurs, the citizen should have an accessible and effective procedure to vindicate their rights. Establishing a functioning system where privacy and data protection are protected thus assists in achieving a free democratic communicative order.[12]

A culture concentrated on protecting privacy within Europe focuses on pluralism, democracy and autonomy. Autonomy is central to conceptualising both privacy and data protection. Differences are to be observed between substantive and informational privacy.[13] Substantive protection allows the individual to engage in daily affairs free from the threat of state coercion or harm. Privacy creates the environment through which informational autonomy can be exercised.[14] Data protection mechanisms such as data portability, rectification and erasure hand the individual greater control over content personal information. In the absence of such controls, human vulnerability increases. As Feldman notes, '[i]f people are able to release [private] information with impunity, it might have the effect of illegitimately constraining a person's choice as to his or her private behaviour, interfering in a major way with his or her autonomy.[15] Effective data protection provisions assist citizens to achieve human development and flourish within society. This ultimately contributes to the maintenance of healthy democracy and encourages civic engagement.[16] Data protection lessens surveillance woes and the feeling of living within a panoptical society.[17] Gutwirth notes that privacy forms a bedrock of democratic society 'because it affects individual self-determination; the autonomy of relationships; behavioural independence; existential choices and the development of one's self; spiritual peace of mind and the ability to resist power and behavioural manipulation'.[18] Data protection mechanisms lead citizens towards actualising greater personal freedom. Data protection tools such as right to be informed, access, rectification, erasure and object place constraints on information monopolists and states' storage of personal data. Such data protection

[11] Kirsty Hughes, 'The Social Value of Privacy' In Beate Roessler and Dorota Mokrosinska (eds), *Social Dimensions of Privacy: Interdisciplinary Perspectives* (Cambridge University Press 2015) 228.

[12] ibid.

[13] Helen Fenwick, *Civil Liberties and Human Rights* (Routledge 2009) 805.

[14] ibid.

[15] David Feldman 'Secrecy, dignity or autonomy? Views of privacy as a civil liberty' (1994) 47(2) Current Legal Problems 42, 54.

[16] Rachel L. Finn, David Wright and Michael Friedewald, 'Seven Types of Privacy' in Serge Gutwirth, Ronald Leenes, Paul de Hert, Yves Poullet (eds) *European Data Protection: Coming of Age* (Springer Science and Business Media 2013) 9.

[17] John Edward Campbell and Matt Carlson, 'Panopticon.com: Online Surveillance and the Commodification of Privacy' (2002) 46(4) Journal of Broadcasting & Electronic Media 586; Malcolm White, 'Bentham and the panopticon: totalitarian or utilitarian?' (1995) 2 UCL Jurisprudence Review 67; David Lyon, 'Everyday surveillance: Personal data and social classifications' (2002) 5(2) Information, Communication & Society 242.

[18] Serge Gutwirth, *Privacy and the information age* (Rowman & Littlefield Publishers 2002) 30.

tools subject states and corporates to greater scrutiny, promote a culture focused on civil liberties and prevent the fuelling of 'surveillance focused' societies.[19] Data protection therefore has a positive effect on substantive privacy. Such mechanisms increase societal well-being overall and provide valuable tools through which the individual can remedy the asymmetric relationship between the citizen and state, where appropriate.

3. Discourse

As outlined, the concepts of privacy and data protection provide clear benefits to society. At a judicial level, these concepts have been conflated resulting in discordance in the adjudication of privacy issues. This Section shall accordingly proceed to consider this conceptual conflation in light of the perspectives of a) Paul de Hert and Serge Gutwirth, b) Lee Bygrave, and c) Orla Lynksey.

A. de Hert and Gutwirth

Paul de Hert and Serge Gutwirth's framework considers privacy to be a 'tool of opacity' and data protection a 'tool of transparency'.[20] This separatist model asserts that privacy and data protection undertake distinct, fundamental functions but, at the same time, remain complementary. Under this model, privacy serves an 'opacity function' by preventing interference into private life, limiting state power and disproportionate encroachment upon the private sphere. Data protection serves a 'transparency function' by defining the rules that make the processing of data permissible. Data protection channels and controls the processing of information, placing obligations on the controller and granting rights to the data subject.[21] This framework is placed against the backdrop of the democratic constitutional state. The complementary but distinct roles of privacy and data protection serve as constraints on state power.[22] According to de Hert and Gutwirth, data protection is a 'catch-all term' for a multiplicity of ideas relating to the processing of personal data.[23] It is through the application of these ideas that governments attempt to reconcile privacy with surveillance, taxation, and the free flow of information.[24]

[19] In exploring the benefits of privacy as a personality right and its contribution to human flourishing, see: Bart van der Stroot, 'Privacy as Personality Right: Why the ECtHR's Focus on Ulterior Interests Might Prove Indispensable in the Age of "Big Data"' (2015) 31(80) Utrecht Journal of International and European Law 25.

[20] Paul de Hert and Serge Gutwirth, 'Privacy, Data Protection and Law Enforcement. Opacity of the Individual and Transparency of Power', in A. Duff and S. Gutwirth (eds), *Privacy and the criminal law*, (1st edn, Intersentia 2006).

[21] ibid at 4.2.

[22] ibid.

[23] Paul de Hert and Serge Gutwirth, 'Data Protection in the Case Law of Strasbourg and Luxenbourg: Constitutionalisation in Action', *Reinventing Data Protection?* (Springer 2009).

[24] ibid.

Providing helpful analysis of de Hert and Gutwirth's framework, Noberto Nuno Gomez de Andrade notes that the tools of opacity and transparency do not exclude each other.[25] On the contrary, 'each tool supplements and pre-supposes the other'.[26] Privacy is a substantive right, while data protection is procedural.[27] Privacy serves as a normative tool, assisting in the realisation of individual freedoms—for example voting for local and national political representation or referendum by secret ballot. The substantive nature of privacy is observed in the judicial exercise of shielding the individual from disproportionate intrusion into one's private life by private and state entities.[28] Procedural rights appear at a later stage, once substantive rights have been weighed, and are formal in design.[29] Procedural rights (such as a right to rectification) aim to hold those who possess power to account. These 'transparency tools' assist in realising the substantive rights environment. Procedural rights formulate the legal conditions and procedures through which substantive rights are expanded.[30] Effective realisation and enforcement of privacy rights are supplemented by data protection rules. The rights which the General Data Protection Regulation creates assist in bolstering privacy rights across the Union. Under this scheme, data protection ranks below privacy and serves a supportive and ancillary function. It places a structure through which the processing of information concerning the individual is respected. As a procedural 'tool of transparency', data protection has no real value; it merely serves as a facilitator of privacy.

De Hert and Gutwirth's analysis is helpful in its simplicity and coherence. The theory properly locates the function of data protection within the democratic constitutional order. Privacy and data protection assist in permitting the individual to maintain control over individuality. Data protection is not a later 'spin off' of privacy, but clarifies the conditions through which processing of information concerning the individual becomes legitimate.[31] This theory illustrates that, given their diverging functions; privacy and data protection are best not placed within the same bottle.[32] De Hert and Gutwirth conclude that 'data protection principles might seem less substantive and more procedural compared to other rights… they are in reality closely tied to substantive values and protect a broad scale of fundamental values'.[33]

[25] Norberto Nuno Gomes de Andrade, 'Data Protection, Privacy and Identity: Distinguishing Concepts and Articulating Rights' in Simone Fischer-Hübner and others (eds), *Privacy and Identity Management for Life*, (1st edn, Springer Berlin Heidelberg 2011) 96.

[26] ibid.

[27] ibid.

[28] ibid.

[29] ibid.

[30] ibid.

[31] ibid 96.

[32] ibid 97.

[33] Federico Ferretti, *EU Competition Law, the Consumer Interest and Data Protection: The Exchange of Consumer Information in the Retail Financial Sector* (Springer 2014) 105.

Despite the apparent clarity of de Hert and Gutwirth's theory, criticisms emerge. Tzanou points out this separation attempts to show the independent value of data protection.[34] However, under this framework, data protection will always be dependent upon and ultimately collapse into privacy. Data protection is denigrated to rules detailing requirements for consent and legitimate processing of information. As a 'transparency tool', data protection rules serve to assist in the realisation of privacy. Even though de Hert and Gutwirth note the benefits of Article 8 of the EU Charter conferring independent constitutional status upon data protection, their formulation requires personal data breaches to be adjudicated with privacy, as opposed to a distinct consideration of data protection, taking precedence. The formulation undermines developing data protection as a distinct fundamental right. It dilutes the significance of informational privacy and data protection rights. Contemporary threats to informational privacy, such as those illustrated by the Edward Snowden revelations, highlight the importance of developing effective legal frameworks. Ensuring data protection and informational privacy are distinct fundamental rights increases protection against such threats.

The authors conclude '[o]pacity and transparency each have their own role to play. They are not communicating vessels.'[35] De Hert and Gutwirth fail to clarify the exact scope of 'opacity'.[36] There are clear benefits to a structure that separates privacy and data protection through the prisms of 'opacity' and 'transparency'. It creates a welcomed separation of the two tools' contrasting core roles. A problem with the theory is its failure to expand upon situations where tools of transparency fall upon opacity to justify their execution and existence. There is a failure to properly consider the opacity dimensions of data protection. This formulation envisages two lines that fail to communicate effectively. A failure to examine the opacity dimensions similar to both, risks the development and value of data protection as a distinct fundamental right. Since there will be areas of overlap and similarity, such a separatist formula raises concern. De Hert and Gutwirth

[34] Maria Tzanou, 'Data Protection as a Fundamental Right next to Privacy? "Reconstructing" a Not so New Right' (2013) 3 International Data Privacy Law 88.

[35] ibid.

[36] In particular, the authors do not define why 'opacity' as a key term is used: 'Opacity tools set limits to the interference of the power with the individuals' autonomy and as such, they have a strong normative nature. The regime they install is that of a principled prescription: they foresee 'no but…' law. Through these tools, the (constitutional) legislator takes the place of the individual as the prime arbiter of desirable or undesirable acts that infringe on liberty, autonomy and identity-building: some actors are considered unlawful even if the individual consents.' Serge Gutwirth and Paul de Hert, 'Regulating profiling in a democratic constitutional state' in Mireille Hildebrandt and Serge Gutwirth (eds), 'Profiling the European Citizen' (Springer 2009).

state that transparency and opacity will blend, but fail to comprehensively work through the opacity/transparency divide, stating:

> [D]ata protection principles might seem less substantive and more procedural compared to other rights norms but they are in reality closely tied to substantial values and protect a broad scale of fundamental values other than privacy. [37]

De Hert and Gutwirth recognise that placing data protection under the ambit of privacy could inhibit the societal benefits of data protection rights. By placing data protection under a purist transparency formula, it neglects the overall value and societal impact of data protection. By failing to effectively outline the relationship of opacity within data protection, it views data protection as an overly procedural mechanism. This results in its ultimate societal value being lost. It places data protection within an unrealistic cocoon. Privacy and data protection in the majority run in different directions. A complete conflation of the two risks conflicts but a complete separation also raises concerns.

1. The Danger of Proceduralisation

De Hert and Gutwirth discuss the 'danger of proceduralisation' in relation to the ECtHR's expansion of the right to respect for private life as encompassing data protection.[38] According to de Hert and Gutwirth, Article 8 ECHR jurisprudence has gone too far in expanding privacy to encompass data protection. Their theory illustrates the problems that can arise when privacy and data protection converge, without any sustained discussion as to their inherent differences. De Hert and Gutwirth believe that Article 8 is no place for procedural developments.[39] Procedural requirements are best located within the Article 13 right to an effective remedy for violations of Convention rights.[40] De Hert and Gutwirth contend that the ECtHR has utilised procedural rights to construe substantive norms. This has led to interpreting 'procedural rights narrowly'.[41] The benefits of proceduralisation include objectivity and impartiality. However, risks include 'the formalization, bureaucratization and de-politicisation of human rights questions'[42] In the well-known phone-tapping case of *Klass v Germany*, for example, the ECtHR outlined in detail the procedural constraints to be complied with for tapping to be legitimate.[43] De Hert and Gutwirth believe the 'necessary in a democratic society' element is

[37] De Hert and Gutwirth (n 23) 44.
[38] De Hert and Gutwirth (n 20) 87.
[39] ibid.
[40] ibid 88.
[41] ibid.
[42] ibid.
[43] *Klass v Germany* (1978) 2 EHRR 214.

neglected. In somewhat extreme sentiment, they assert proceduralisation 'might well bring the erosion of recognized rights'.[44]

If we envisage Article 8 as primarily dealing with protecting zones of opacity, bringing data protection elements under its remit is beyond its scope. Following this reasoning, questions surface concerning the ECtHR's future legitimacy in dealing with privacy claims. If the Court wished to develop data protection freedoms under Article 8, it needs to be clear as to the exact value of privacy in its foundation. Kirsty Hughes has recently outlined the societal values inherent to privacy within the ECHR and the ECtHR's overall principles.[45] Hughes submits that there is a failure of the ECtHR to articulate the societal value of privacy. To bolster the intellectual consistency of the Court, it should recognize that the value of privacy is essential to the democratic state and 'is crucial to facilitating harmonious social interaction'.[46] The ECtHR concentrates on the legality requirement and limits the 'necessary in a democratic society' discussion. Once the Court addresses if there is a legal basis for the infringement and finds a breach, it does not address the issue as to whether it the measure is 'necessary in a democratic society'. There is no sustained discussion as to how data protection freedoms contribute toward and are necessary to democratic society. In setting the foundation from which the ECtHR draws inspiration and adjudicates cases, it would require more from the state to justify proportionate interference.[47] An effective data protection framework should strive to locate the value and inspiration of data protection. As the ECHR remains ambiguous as to the central value of privacy, the Court has incorporated procedural elements. By failing to properly locate the conceptual foundations of privacy, confusion is increased when the Court strays into areas not traditionally envisaged by the Convention, like data protection. If the Court wishes to work through data protection through the ambit of Article 8, it should be clear from first principles as to the exact relationship between the two. Problems inherent to internal expansion notwithstanding, given that the Convention is a 'living instrument',[48] one can see a need to interpret Article 8 expansively to bring informational privacy under its ambit and Article 8 is arguably broad enough to found such development. Even though it may be misplaced, such development is necessary.

[44] De Hert and Gutwirth (n 20) 89. The authors ultimately feel that transparency mechanisms have no place within Article 8, concluding that the drafters of the Convention could not have envisaged the development of Article 8 as a source of procedural conditions and rights. This forms a view that Article 8 jurisprudence on this issue is misplaced and not currently fit for purpose.

[45] Kirsty Hughes, 'The Social Value of Privacy' (n 11) 228.

[46] ibid 238.

[47] ibid 240.

[48] George Letsas, 'Strasbourg's Interpretive Ethic: Lessons for the International Lawyer' (2010) European Journal of International Law 509.

2. The Charter

We observe De Hert and Gutwirth's opacity/transparency divide clearly within Article 7 and Article 8 of the EU Charter. Article 7 of the EU Charter unimaginatively recites Article 8 ECHR stating that 'Everyone has the right to respect for his or her private and family life, home and communications'.[49] Article 7 of the EU Charter falls under the 'opacity' formulation, preventing unwarranted intrusion by the state within the private sphere. Article 8 of the EU Charter outlines the 'transparency' elements supporting the realisation of Article 7's substantive formulation, guaranteeing the 'right to protection of personal data concerning him or her'[50] and requiring data be processed fairly and on the basis of consent or some other legitimate basis laid down by the law.[51]

While Article 7 EU Charter has its influence in Article 8 ECHR, given there is no distinct right to data protection under the ECHR and the piecemeal fashion through which the ECtHR has developed data protection freedoms, Article 7 EU Charter may be problematic when assessing its precise relationship to Article 8. The manner in which the CJEU adjudicates Article 8 data protection claims illustrates a continued reliance on the Article 7 right to privacy.[52] Currently, the two rights continue to be conflated with no clear conceptual footing. De Hert and Gutwirth's theory nonetheless illustrates the difference in logic between Articles 7 and 8 of the EU Charter. The underlying rationale for the separation of the two within the Charter remains unclear, perhaps purposely so. Their framework helps inform a broader picture, yet fundamentally neglects how the two should effectively communicate with and complement each other.

B. Lee Bygrave

In response to de Hert and Gutwirth's failure to locate the place of 'opacity' within data protection, Lee Bygrave's work contributes to examining this perceive gap. Bygrave expands upon the 'opacity' nature of data protection.[53] Bygrave asserts that it is not effective for data protection to be characterised or centrally concerned with privacy. It is under this construction that data protection is about the reconciliation of the interests of the data subjects with the legitimate interests of data controllers.[54]

[49] The wording of Article 8(1) is almost identical stating that 'Everyone has the right to respect for his private and family life, his home and correspondence'.

[50] Article 8(1), EU Charter.

[51] Article 8(2), EU Charter.

[52] Felix Bieker, 'The Court of Justice of the European Union, Data Retention and the Right to Data Protection and Privacy—Where Are We Now?' in Camenisch, Fischer-Hubner and Hansen (eds), *Privacy and Identity Management for the Future Internet in the Age of Globalisation* (Springer 2015).

[53] Lee Bygrave, 'The place of privacy in data protection law' (2001) University of New South Wales Law Journal 241, 277–283.

[54] ibid.

Even though the evolution of data protection has been somewhat convoluted, Bygrave details nine elements of data protection: 'the fair and lawful processing principle, the transparency principle, the data subject participation and control principle, the purpose limitation principle, the data minimization principle, the information quality principle, the proportionality principle, the security principle, and the sensitivity principle.'[55] A failure to characterise privacy in terms of data protection is reflected in the difficulties in attempting 'to give privacy a precise, analytically serviceable and generally accepted meaning.'[56] The law's guidance and development is affected by the current dysfunctional relationship between the two. Bygrave is correct that the expansive nature of privacy forms part of the rationale for data protection. [57]

In line with the increased emphasis within the Article 8 ECHR jurisprudence considering the of benefits privacy for society and democracy, the values of personal identity that data protection helps to realise 'have a broader societal significance.'[58] In bridging the gap between privacy and data protection, Bygrave is correct in asserting that the general societal values common to both must be recognised when forming effective legal policy on data protection issues. De Hert and Gutwirth's theory neglects the impact that procedural data protection structures can have in modern constitutional democracy. It does so by failing to acknowledge the opacity and substantive benefits data protection has to the citizen and society. Second, data protection is concerned with 'setting standards for the quality of personal information' which has 'little direct connection' to privacy values.[59] Third, data protection rules are concerned with the legitimate processing of information, taking on a management quality.[60] In this sense, data protection orbits privacy, but never seeks to be attached. Both concepts coexist to prevent conflict by putting in place appropriate management strategies.[61] Bygrave concludes that while privacy occupies a place within data protection, viewing data protection as serving and securing privacy is problematic; data protection serves a multitude of interests that extend beyond privacy.

The focus of data protection laws shifts with technological developments. Attempts to create a 'right to be forgotten' illustrate an example of recent policy adapting to behavioural change.[62] The emergence of European data protection is observed within Convention 108, the 1995 Directive, and, most recently, the Regulation. The Regulation, due to come into effect in the spring of 2018, is a

[55] Frederik Zuiderveen Borgesius 'Privacy on the Internet' in Alberto Alemanno and Anne-Lise Sibony (eds) *Nudge and the Law: A European Perspective* (Bloomsbury Publishing 2015) 183.
[56] Bygrave (n 53) 278.
[57] ibid 281.
[58] ibid.
[59] ibid.
[60] ibid 282.
[61] ibid.
[62] C-131/12 *Google Spain* (n 2).

significant step in the evolution of European data protection policy. It contains a number of innovations, such as introducing a right to data portability and a right of erasure. As privacy concerns continue to increase with changes in human behaviour, so too does the focus of data protection laws and regulation. Yet we should still strive to locate the guiding principles and conceptual structures of informational privacy in order to bolster the intellectual consistency and future of this area. Collapsing data protection into privacy effectively renders the essential societal and democratic benefits of such rights incoherent and haphazard. The risk of the Article 8 ECHR approach where the ECtHR views data protection as a mere subset of privacy, risks diluting the constitutional value of data protection to 'soft law'.[63]

C. Orla Lynskey

Lynskey believes while data protection and privacy overlap, data protection offers individuals more rights than privacy. Data protection more effectively ensures 'selective presentation' of individual identity than the right to privacy, 'thereby promoting self-development and the personality rights of the individual.'[64] It assists in identity construction. By providing individuals with greater control over their personal data, the individual can 'reveal different elements of their personality.'[65] Effective data protection frameworks thus focus more on controlling disclosure of information. Privacy is not as focused in its informational management function. Data protection tools are concentrated on removing power and information asymmetries in the relationship between the individual and the data controller/processor.[66] Power asymmetries impact the ability of the individual to make an informed choice about whether to allow their information to be processed or not.[67] The right to data protection goes further than the right to privacy as it envisages that 'individuals . . . have difficulty asserting their preferences for privacy protection'.[68] Effective rules help to empower citizens and reduce power asymmetries. The right to data protection thus hands the individual more control than the right to privacy. Lynskey advises that the continued conflation of the right to privacy and the right to data protection is best avoided. Recognising the right to data protection as a right distinct from the right to privacy re-balances the asymmetric relationship between the individual and state or private entities. As the 'cascade of decaying information'[69] concerning the individuals online information only sets to increase, such protections have clear significance.

[63] De Hert and Gutwirth (n 23) 44.

[64] Orla Lynskey, 'Deconstructing Data Protection: The "Added-Value" of a Right to Data Protection in the EU Legal Order' (2015) 63 International & Comparative Law Quarterly 590.

[65] ibid 591.

[66] ibid 592.

[67] ibid.

[68] ibid 594.

[69] Oskar Josef Gstrein, 'The Cascade of Decaying Information: Putting the "Right to Be Forgotten" in Perspective' (2015) 21 Computer and Telecommunications Law Review 40.

Lynskey's work is helpful in detailing the value of having a clear distinct right to data protection within the EU constitutional order. Lynskey explains that the CJEU continues to conflate the two rights in its adjudication after the Lisbon Treaty.[70] The discourse should focus on how best to ensure that the right data protection and the right to privacy should complement and inform, rather than converge. This approach will assist in further bolstering data protection rights for EU citizens. The earlier discussion about the 'dangers of proceduralisation' illustrates that a continued reliance by the CJEU on Article 8 ECHR to interpret the Data Protection Directive may be problematic. To ensure data is processed lawfully, the CJEU can rely solely on Article 8 of the EU Charter without resorting to reliance on Article 8 ECHR. Nonetheless, it is questionable whether strict reliance on Article EU Charter alone will be effective.

4. Comment

This brief review and assessment of the theories of de Hert and Gutwirth, Bygrave, and Lynskey illustrates the ambiguity surrounding the conceptual bases of privacy and data protection under the EU Charter and ECHR. Such ambiguity impacts upon adjudication of these issues within the Courts. This ultimately amounts to a disservice to expanding EU privacy rights. The three theories discussed each complement one another in informing a broader perspective. De Hert and Gutwirth's opacity-transparency formulation provides a clear framework for understanding the functional differences between the privacy and data protection. The theory fails, however, to discuss areas of overlap between privacy and data protection and fails to encourage effective dialogue between the two. It fails to discuss the 'opacity' nature of data protection and how it should engage with competing rights. Lee Bygrave's theory assists in these communication failures in correctly identifying that data protection assists the actualisation of privacy rights. Data protections aims and values go beyond privacy, however. Orla Lynskey finally places this continued conflation within its current context within the CJEU. Lynskey correctly identifies the importance in recognising data protection as a distinct right in increasing access to informational self-determination and preventing power asymmetries. In order for this potential tension between privacy and data protection to ease, we need to be clear about the exact values of data protection to society and democracy. Similar to the manner in which the ECtHR has failed to expand upon the benefits of privacy for societal well-being, failing to concentrate on the overall value of data protection mechanisms for the individual's benefit within democracy will decrease rather than bolster data protection rights.

Data protection is located on the fringes of privacy and many of the values and justifications for data protection overlap with those for privacy. Data protection

[70] Orla Lynskey, 'Deconstructing Data Protection: The "Added-Value" of a Right to Data Protection in the EU Legal Order' (2015) 63 International & Comparative Law Quarterly 569, 579–581.

deduces its foundational inspirations from privacy, but remains distinct. Emphasis should be placed on data protection beyond a purely procedural or mechanical function. Effective data protection rules empower citizens and hand citizens greater control over personal information. This assists in achieving a free democratic order that values the individual's private sphere. Data protection is not of solely structural significance; it provides the structures through which the private sphere is respected. Whilst data protection rules outline principles that result in the proper processing of information, such structures have significant impact to the individual and democratic society. The two rights are reciprocal.

The current judicial approaches to Articles 7 and 8 of the EU Charter and Article 8 ECHR are inconsistent. Such inconsistency stems in part from a reluctance of the courts to clearly articulate the value of privacy to society. There is limited intellectual stability from the ECtHR in its development of the right to informational privacy. This flows from the ECtHR's inability to locate its intellectual footing in relation to the Article 8 ECtHR's substantive right to privacy. The CJEU's development of data protection freedoms under Article 8 of the EU Charter is similarly unstable. We remain unclear about the relationship between the Article 7 EU Charter right to privacy and the Article 8 EU Charter right to data protection, its impact upon private entities, and the place of Article 8 ECHR jurisprudence within the EU Charter framework. Increased discussion at a judicial level in locating the conceptual place of informational privacy assists in increasing legitimacy of the CJEU's recent emboldened approach. de Hert and Gutwirth's 'danger of proceduralisation' comment illustrates the problems of failing to give data protection independent constitutional status. It demonstrates the failure of the ECtHR to satisfactorily locate the intellectual routes of privacy.

In *Digital Rights Ireland*,[71] the CJEU utilised its own cases and Article 7 of the EU Charter to find the retention of data was unlawful.[72] It then referred to ECtHR cases to find that access to the data was a separate interference.[73] In seeking to outline the differences between data protection and privacy under Article 7 and Article 8 of the EU Charter, it ultimately enunciated identical versions.[74] *Google Spain*[75] epitomises the minimalist nature of the CJEU's reasoning and the problems this creates for the actualisation of the CJEU's judicial innovations. The CJEU gave no detailed assessment of the right to privacy or the competing considerations at play. In stark contrast to *Schrems,* there was a lack of engagement

[71] *Digital Rights Ireland* Joined Cases C-293/12 C-594/12 8 April 2014 (n 3).

[72] ibid paras. 33–34.

[73] ibid para. 35.

[74] Felix Bieker, 'The Court of Justice of the European Union, Data Retention and the Right to Data Protection and Privacy—Where Are We Now?' in Camenisch, Fischer-Hubner and Hansen (eds), Privacy and Identity Management for the Future Internet in the Age of Globalisation (Springer 2015).

[75] *Google Spain* (n 2).

with ECtHR judgments considering the Internet archives.[76] In determining when the 'preponderant interest'[77] where public interest requires retention of the material online, reference to *Von Hannover (no. 2)*[78] which considered such considerations in detail, would have been welcomed. *Google Spain* illustrates a failure of the CJEU to clarify the opacity dimensions of data protection rights, looking beyond the delisting of the search result as a procedural tool but how actualized the privacy rights engaged. *Schrems* relied upon Article 7 EU Charter in holding the intrusion violated the 'essence of privacy'.[79] It was not satisfactorily expanded upon what constitutes the 'essence of privacy', especially with reference to Article 7 EU Charter and what role (if any) Article 8 EU Charter plays within such adjudication.

In examining these judgments, it is difficult to gauge any broader or consistent reading as to the CJEU's conception of privacy and data protection. A contributory factor to this judicial inadequacy is a fundamental failure to articulate the foundations of data protection and privacy to the individual and society. The EU Charter paves the way forward at a European level for the protection of informational privacy. The CJEU's inability to effectively interpret and expand its provisions renders such judicial innovations haphazard and undermines the legitimacy of the CJEU in developing data protection rights. EU data protection can be viewed as both a sword and a shield. It protects individual's information in daily life from unlawful intrusion by states and private entities. Recent cases from the CJEU illustrate a judicial institution not afraid to utilise such provisions to hold both states and private entities to account for violations. This is so having little regard to possible political critique or economic considerations. On the other hand, within the ECHR, we may have to reassess whether Article 8 ECHR (in its current form) is the best place for expanding data protection freedoms. The Article 13 ECHR right to an effective remedy may be an appropriate alternative. It is not possible for data protection rights to be effectively actualised when the ECtHR remains unable to outline the intellectual stability of Article 8 ECHR right to privacy. Similarly, we remain unclear as to the foundations of the right to privacy and the right to data protection within the CJEU, the relationship between Article 7 and 8 of the EU Charter, and the place of Article 8 ECHR jurisprudence. In tackling these issues, both the ECtHR and the CJEU need to be clear from first principles about the relative significance of informational privacy. A failure to

[76] See *Case of Times Newspapers Ltd. (Nos. 1 and 2) v. The United Kingdom* Applications nos. 3002/03 and 23676/03, 10 March 2009; *Editorial Board of Pravoye Delo and Shtekel v. Ukraine* Application no. 33914/95, 5 May 2011; *Ahmet Yildirim v. Turkey* 18 December 2012, application No. 3111/10; *Wegrzynowski and Smolczewski v. Poland* Application No. 33846/07, 16 July 2013.

[77] *Google Spain* (n 2) para 98.

[78] Applications nos. 40660/08 and 60641/08, 7 February 2012.

[79] *Schrems* (n 4); see Martin Scheininm 'The Essence of Privacy, and Varying Degrees of Intrusion' *Verfassungsblog,* 18 November 2015 Available at: http://www.verfassungsblog.de/en/the-essence-of-privacy-and-varying-degrees-of-intrusion/#.Vkxi3nvFk14.

properly locate the place of data protection within Article 8 ECHR jurisprudence is problematic. Moreover, a systematic failure of the CJEU to work through the relationship between data protection and privacy and the possible competing considerations (most fundamentally, freedom of expression), raises questions as to the effectiveness of the Court's recent emboldened stance.

5. CONCLUSION

What role does the right to data protection play with regard to the right to privacy? Clearly the right to data protection and the right to privacy are interrelated and often overlap. Both add something of clear significance in providing for a free democratic communicative order. Data protection in the majority goes beyond privacy, but the two should communicate effectively. Recognising that data protection is a distinct right is necessary to ensure its continued expansion as 'hard law' at EU constitutional level. We must also recognise that from a conceptual perspective, a complete separation is not possible. This question comes at a time of increased discussions surrounding the Regulation. Whatever direction informational privacy may go, it should be guided with transparency in mind and preventing unnecessary tension within the two frameworks.

There are two principal conclusions. First, de Hert & Gutwirth's theory acts as a valued basis for assessing the relationship between the two. Their theory fails, however, to examine the 'opacity' dimensions of data protection and how it can effectively deal with competing rights. It does identify a procedural or 'transparency' line that mirrors the ECHR and EU Charter frameworks. Bygrave fills in some of the gaps in de Hert and Gutwirth's formulation. According to Bygrave, data protection orbits privacy, but ultimately remains a distinct right. Lynskey's work recognises the importance of viewing data protection as a distinct right in reducing power asymmetries and promoting informational self-determination. Moreover, she illustrates the importance of transferring discussion of this conceptual conflict to a judicial level. The ECtHR should reassess the exact place of data protection. The ECtHR should, to borrow a phrase from *Schrems*, reconsider what is the 'essence of privacy' and what role it plays within society. Only then can we assess the respective place of data protection.

Whatever problems the ECtHR has in terms of conceptualising privacy, the CJEU is a cause for greater concern. Fundamental rights may be viewed as a new area for the CJEU. If the CJEU continues its fervent expansion of privacy rights, it needs to be clear about its conception of privacy and the distinct but complementary roles of the EU Charter and ECHR. In order for such judicial growth to gather legitimacy, clarity concerning the conceptual roles of privacy and data protection is necessary. Otherwise, pivotal judicial developments created by litigants such as Max Schrems, Mario Costeja González, and Digital Rights Ireland could be futile.

The Service Conception and Normative Collective Action

Guy Ziv-Shalom[1]

1. Introduction

Joseph Raz's service conception of authority[2] is the most influential account of normative authority available today. It capably describes and explains the 'right to rule'—understood as the power to impose duties to obey upon subjects—as fulfilling the role of servicing the governed. However, its straightforwardness and flexibility also pose a potential obstacle. It has been argued that the criteria for legitimate authority as specified in the service conception are unfitting, as they allow directives to bind subjects where it is clear that no authority exists, like in the case of financial advice.[3] It has also been argued that it cannot explain parental authority[4] or the authority of criminal law.[5] The goal of this article is to add further details and observations to the service conception that may also help to clear up possible misapprehensions in regard to the theory, showing that some of these objections are misplaced.

The main point is that, according to the service conception, when there is no independent, prior obligation to obey, authority is only legitimate when it helps in solving certain problems that I will refer to as 'normative collective action problems'. These are familiar types of problems such as coordination problems and prisoners' dilemmas, which hamper the agent's ability to conform to reason. The involvement of individuals other than the agent creates a situation in which

[1] LL.M. Candidate, Hebrew University of Jerusalem Law Faculty. I am deeply indebted to David Enoch for his guidance. I also thank Re'em Segev, Scott Hershovitz, Sandra Ziv and the editors of the *Cambridge Law Review*.

[2] Joseph Raz, 'The Problem of Authority: Revisiting the Service Conception' (2006) 90 Minn L Rev 1003.

[3] See, for example, Stephen Darwall, 'Authority and Reasons: Exclusionary and Second-Personal' (2010) 120 Ethics 257 (2010).

[4] Scott Hershovitz, 'The Role of Authority' (2011) 11 (7) the Imprint 1, 11–12.

[5] See Scott Hershovitz, *Accountability and Political Authority* (unpublished draft, <http://sites.google.com/site/nicosstavropoulos/Hershovitz_AccountabilityandPolitica.pdf>).

acting in accordance with rational capacity alone may lead to actions that would not be rationally justified, were the agent capable of relying on the behaviour of others. The purpose of authority, according to Raz, is to enable us to 'secure preexisting goals in ways not otherwise possible.'[6] It is the very issuing of the directive that expands our capacity for rational action, overcoming the natural obstacles that stem from living in a society where brushing up against each other is unavoidable. It opens up new possibilities, hitherto unavailable to conform better to reason. Relying on the behaviour of others is key if a more rational action is to be taken in such cases, and it is only this need—when compliance is central for the achievement of our own goals as individuals in a society—that can provide the moral justification for substituting one's own judgment with that of the authority. This also means that, in such cases, de facto authority—descriptively, successfully exercising power over subjects—is a precondition of legitimate authority. In this respect, this interpretation of the service conception further specifies the circumstances in which legitimate authority can be said to be established.

It is important to stress that the term 'collective action problems' is employed here in a broad sense, beyond its regular use in the game-theoretical economic context. In its usual meaning, actors have the sole objective of maximising their self-interest. But the collective action problems I refer to—those relevant for practical authority—are not particularly those in which the agent is trying to act in her own best interest. Normatively, agents have reasons that apply to them regardless of their self-interest, such as reasons to keep promises or to respect one another. In other words, '[T]he coordinated schemes of action that political authorities should pursue are those to which people should be committed, or those needed to secure goals that people should have, which are not always the goals which they do have'.[7] John Finnis similarly observed that

The first difference, then, between the concept of co-ordination problem used in game theory and the concept appropriate for political or legal philosophy is that the latter must extend to include situations where, in relation to the 'situation' itself and the interests of the parties in that situation as such, there is no convergence or sharing of interests. And the second difference will be that political and legal theory must take into consideration a type of 'interest' systematically excluded from game theory (and whose exclusion is particularly evident in the game-theoretical handling of Prisoners' Dilemma problems), viz. interest in the fairness of the game's play and outcome, which any player can prefer to an increment in the advancement or protection of his 'own' interests.[8]

The term is therefore borrowed to reflect the fact that acting rationally without some type of coordination would not be the best way of conforming to our normative reasons. This demands that the agent take all of the relevant reasons for

[6] The Problem of Authority, (n 2), 1034.
[7] ibid 1032.
[8] John Finnis, 'Law as Co-ordination' (1989) 2 (1) Rat Jur 97, 100.

action into consideration,[9] not just those that concern her self-interest. Hence the addition of the word 'normative' to the otherwise familiar term.

It seems that Raz himself remains vague on the precise nature of the connection between collective action problems and the service conception. In several places, Raz acknowledges that securing coordination is a 'major, if not the main factor in establishing the legitimacy of political authorities',[10] and he also mentions the disentangling of prisoners' dilemmas in this context.[11] However, Raz avoids describing this connection in detail or explaining the difference between parental and political authorities in this respect. This article will tackle these issues and attempt to outline the connection between collective action problems and legitimate authority.

In Part A of this Section, I will briefly present the service conception. In Part B, I shall then utilise Enoch's observations regarding authority as a particular case of robust reason-giving, which is essential if we are to understand what sort of cases are relevant for the Normal Justification Thesis. Thereafter, Section 2 of this Article will apply the previous Sections and Section 3 will address possible objections, arguing that if Raz and Enoch are correct—and assuming it is not one of the instances where there is no prior, independent obligation to obey—practical authority is only legitimate when it solves normative collective action problems: problems pertinent to complex societies, which hamper the agent's ability to conform to the reasons that apply to her.

A. The Service Conception, in a Nutshell

The service conception addresses the paradoxical nature of authority. Assuming that our concept of authority entails the capacity to manufacture duties for subjects out of thin air, requiring one to abandon autonomous discretion, an account of authority must address the troubling moral problem of 'how can it be consistent with one's standing as a person to be subject to the will of another in the way one is when subject to the authority of another?'[12]

Raz's answer is comprised of several theses. The core thesis, and the focal point of this article, is the Normal Justification Thesis (hereinafter 'NJT'): that 'the normal way to justify authority is that the subject would better conform to reasons

[9] Raz limits the relevant reasons to categorical reasons. These are reasons 'whose application is not conditional on the agent's inclinations or preferences, and so on.' See Joseph Raz, 'On Respect, Authority, and Neutrality: a Response' (2010) 120 Ethics 279, 291.

[10] The Problem of Authority (n 2) 1031.

[11] Joseph Raz, 'Authority and Justification' (1985) 14 (1) Philos Public Aff 3, 17, 21; see, also, On Respect (n 9), 301.

[12] The Problem of Authority (n 2) 1014.

that apply to him anyway, if he intends to be guided by the authority's directives than if he does not.'[13]

The second condition is the independence thesis: that the matters regarding which the NJT is met are such that, with respect to them, it is better to conform to reason than to decide for oneself.[14] This condition requires that, in the respective case, self-judgment will have no such intrinsic value that will outweigh the benefits of yielding to another's discretion. This condition is not the subject of this article.

Raz explains that we should recognise that self-judgment is very much of instrumental value, at least in part. Our rational capacity 'derives from the fact that there are reasons that we should satisfy, and this capacity enables us to do so.'[15] Self-judgment is not, however, the only means to achieve the end of conforming to reason. If yielding to a directive makes us conform to the reasons that apply to us anyway better than we would without it, we now have a reason to abide by it. Authoritative directives, therefore, must enable subjects to conform better to the background reasons that apply to them anyway if these directives are to be legitimate.

This, however, is only part of the story, as legitimate authoritative directives do not simply produce another reason for action, adding to our set of existing background reasons to factor into our deliberations, eventually using our self-judgment to decide what action to perform. After all, the significance of the duty to obey is that we are to carry out the task required without contemplation. We are not to 'second-guess the wisdom or advisability of the authority's directives.'[16] When we receive an authoritative directive, we are to set aside certain other reasons for action—specifically, reasons such that acting on them might lead to a failure to perform the task specified in the directive. This is what sets authority apart—its directives preempt the background reasons that 'might militate against the authoritative directives and replace them with their own requirements,'[17] thus entailing an obligation to obey. Self-judgment is therefore normatively removed through the concept of preemption, or the preemption thesis—the mechanism behind the creation of a duty. But how exactly does preemption take place?

Raz explains this by using the notion of exclusionary reasons. An exclusionary reason is a second-order reason. More specifically, it is a reason not to act for a first-order reason (or reasons). It excludes them. Once a legitimate authoritative directive is issued, it provides us not only with reason for action, but also with an exclusionary reason: a reason not to act for some of the first-order reasons. Particularly, these first-order reasons are reasons which acting on them would result in failure to perform the task required by the directive. These reasons are

[13] ibid.
[14] ibid.
[15] ibid 1017.
[16] ibid 1018.
[17] ibid 1019.

those on the 'losing side of the argument';[18] those which acting on them would not lead us to conform better to reason. (Even though acting on them would satisfy these reasons in particular, conforming to reason overall demands that all relevant reasons that apply to the agent be taken into consideration and assigned the appropriate weight in determining what action will make us conform best to it.) The reasons excluded are only those that the authority has taken into account in its calculation; if an unexpected circumstance arises which the authority has not taken into account, obeying the directive will not necessarily make us conform better to reason, thus rendering the directive unbinding.[19]

Once the NJT is satisfied, it is easy to see how an authoritative directive provides not only a first-order reason to follow the directive, but also an exclusionary reason. The authority considered and weighed our background reasons for us, and gave us an answer that reflects how to act on them. If the NJT is satisfied, following it will enable us to conform better to our own independent background reasons. (Even though it may not always be the optimal answer that could be given). Once we are given such an answer, we have a reason not to act on reasons that may lead us astray, failing to conform to reason in the manner we would otherwise achieve. In Raz's words, 'the mediating role of authority cannot be carried out if its subjects do not guide their actions by its instructions instead of by the reasons on which they are supposed to depend.'[20]

Reasons for action that also act as exclusionary reasons are sometimes referred to as 'protected reasons',[21] and according to Raz, protected reasons amount to obligations. Thus, the service conception explains how authority manufactures duties, or at least protected reasons, out of thin air.

B. Robust Reason-Giving and the NTJ's Theoretical Appeal

Before moving on to Section 2, it is useful to employ David Enoch's analysis in respect to the type of reason-giving that the service conception attempts to explain, as 'reason-giving' can be an equivocal term. This analysis is necessary to clarify the precise nature of authoritative reason-giving.

Enoch distinguishes three senses of reason-giving. The first kind is epistemic reason-giving, which operates on a purely epistemic level. Its 'giving' is in fact *showing* or *indicating* a reason that is already there, as a background reason already applying to the agent.[22] To borrow Enoch's example, suppose I decide to tell a colleague how much I dislike him. You urge me not to do so. My response goes

[18] ibid 1022.

[19] ibid. See, also, Timothy Endicott, 'Interpretation, Jurisdiction, and the Authority of Law' (2007) 6 (2) American Philosophical Association Newsletter on Philosophy and Law 14, 16.

[20] Joseph Raz, 'Authority, Law, and Morality' in Joseph Raz, *Ethics in the Public Domain: Essays in the Morality of Law and Politics* (Oxford: Oxford University Press, 1994) 198–199.

[21] Joseph Raz, 'Reasoning With Rules' (2001) 54 CLP 1, 14.

[22] David Enoch, 'Authority and Reason-Giving' (2014) 89 (2) Philos Phenomen Res 296, 3–4.

something along the lines of 'give me one reason not to!' and you, in turn, reply by noting the negative implications for the intellectual atmosphere in the department. In this example, it seems you have indeed 'given' me a reason not to share with my colleague my opinion of him, but this giving was in the form of showing me a reason that was already there, irrespective of your giving it to me.

The second type of reason-giving is what Enoch terms as *merely-triggering* reason giving. It is different from the former type in how it triggers dormant reasons, which are independent background reasons already applying to the agent. The triggering is achieved by manipulating the non-normative facts; but that is all they do. They merely trigger a dormant reason for action. To use one of Enoch's examples again, if your neighbourhood grocer raised the price of milk, one might say she has given you a reason to reduce your milk consumption; you had had no reason to do so before she raised the prices, and now you do. But notice the particular kind of giving here. You have a general, independent background reason to save money. This reason does not depend on the grocer's actions. The grocer has simply manipulated the relevant non-normative circumstances. This manipulation has indeed given a reason, but this giving has merely triggered acting on a reason that was there all along.

These two types differ from the third type of reason-giving, which represents a more robust sense thereof. This type triggers a background reason and, moreover, it seems to create a new one, in a more robust manner, as in the case of requests and promises. Requesting and promising create a new reason for action that was not there before. It normatively changes the addressee's set of current reasons for action. They not only reflect the existing set of reasons but actually change them, just by communicating the intention of doing so.[23] If I promise my mother to visit her on the weekend, I now have a new reason to visit her that did not exist before the promise. I have triggered a general reason to keep promises, but also created a reason that was not there before, adding it to my current set of reasons.

This latter type of reason-giving is the one relevant for practical authority. This is due to the fact that we perceive the dictate itself—the very communication of it—as providing a new reason for action, which we did not have before. Given our intuitive proclivities regarding its ability to impose new duties by expressing an intention to do so, our concept of authority is a particular instance of robust reason-giving. The consequences of this observation will be elaborated later in this article. It is, however, important to emphasise at this point that, since practical

[23] ibid 5, 15–16.

authority is a particular case of robust-reason giving, any reason-giving that is beside this type will simply render the NJT irrelevant in the first place.[24]

As noted, robust reason-giving still depends on triggering our background reasons, and duties themselves constitute such reasons. For example, agents can have a background duty to obey someone or something. Some believe that there is a duty to obey the dictates of a regime elected with the proper procedure (such as a treating its subjects as free and equal, or something of the sort). Some argue that this serves as a counterexample to the Razian account: if the legitimacy of authority also depends on such inputs as procedural conditions rather than on the output—the substance of the directives themselves—then the service conception is mistaken to ignore this central element. Raz maintains that such cases are not counterexamples; on the contrary: these are cases when the NJT is satisfied almost trivially. An agent's obligation to obey the regime's commands constitutes a reason, and obeying the command is naturally the best way to conform to such reason.[25]

It is clear, however, that this kind of explanation provided by the NJT for authority is somewhat unsatisfactory. Presupposing a prior duty to obey so that the NJT simply 'confirms' authority is just not that interesting. The interesting question of authority is how it can create a new obligation to obey without the need to rely on some preexisting obligation to obey someone or something (one that the service conception does not attempt to determine). In other words, in such cases the NJT is not a very helpful criterion for the determination of an obligation to obey.[26] (I will call this the 'boring' scenario.) This is particularly important in political contexts, in which presupposing a prior obligation to obey authority is problematic from a liberal point of view. But there are cases in which the service conception need not rely on such prior conditions to determine the legitimacy of authority. It is in these instances—those without a prior obligation to obey—that the NJT performs meaningful theoretical work. These are situations involving collective action problems, as will be elucidated in the following Part.

2. Authority as Solving Normative Collective Action Problems

Agents have strong background reasons to drive safely and efficiently, as well as to avoid endangering the lives of others. Drivers will better conform to these reasons if they act according to the directive to drive on the right than if they do not; due to the disorder in the derelict roads, this would probably result in very slow

[24] Some have an altogether different concept of authority. This concept does not assume that authoritative directives necessarily constitute a new reason for action. Rather, they allow for a more lenient sense of reason-giving, with more room for epistemic components in reasons and reason-giving. See, for example, Donald Regan, 'Authority and Value: Reflections on Raz's Morality of Freedom' (1989) 62 So Cal L Rev, 995, 1021. The service conception as I understand it, however, does assume that authority entails reason-giving in the robust sense, and I shall proceed in light of this assumption.

[25] The Problem of Authority (n 2) 1030.

[26] See Enoch (n 22) 19–20.

or dangerous driving. When a directive instructs us as to which side to drive on, its decree is not intended to affect the agent alone. It also addresses the relevant community—other drivers. De facto authority allows for changing non-normative facts in the world—in this example, that everyone should drive on the right. But the issuing does more than merely changing the non-normative facts. By doing so, it also creates new, better options for rational action that were not there before. The driver is now able to predict the behaviour of others with a high degree of certainty, allowing her to drive more safely and efficiently than before; that is, with improved conformance to her background reasons.

Thanks to this improvement, the driver has now been given a reason to drive on the right, as dictated by the command. Since doing otherwise would make the driver conform worse to her background reasons, this is also a reason to avoid referring to reasons that conflict with the command. Like promises and requests, this reason is given in a robust manner, creating a new reason for action that was not there before, by the mere successful communication of the intent to do so. Thus, the driver is robustly given a protected reason to obey.

It is also worth noting that, in such cases, the ability to solve coordination problems depends on the existence of a non-normative factor: de facto authority. The capacity to address a wide array of people who would actually follow the directive and perhaps even use force to guarantee compliance is a prerequisite of the ability to solve coordination problems. The reaction of other drivers to the decree is thus crucial for a reason to be given, since the reason to drive in the right lane only holds if all or at least a substantive majority of other drivers do the same.

This outcome is not limited to coordination problems of this kind. Similarly, another example of this type of situation is prisoners' dilemma cases. Here, too, there is a sort of collective action problem in which individuals cannot achieve desired results without outside interference that would secure certainty. The authoritative directive changes the way others behave, giving the agent a reason to obey as well, since obeying in this new situation—and reaping the fruits of cooperation made possible through the obedience of others—will have the agent conform better to her independent background reasons.[27]

There are other possible instances, beyond problems of the types described above. These are instances in which we have certain prior general obligations, but conforming to them may prove difficult. This is because we live in large, complex societies and it can be unclear how to discharge obligations towards the people constituting a large moral community. These obligations provide us with independent background reasons, because the obligations themselves constitute reasons for action. An authoritative directive, with its de facto power and capacity

[27] Authority and Justification (n 11) 17–18.

to address and be obeyed by a large number of people, can help us to better conform to these reasons.

This is illustrated by the following example. Let us assume a public good, which needs to be funded by the entire community, as they are all beneficiaries. Let us also reasonably assume that I have a general obligation to carry my own weight in the community. It is unclear to me, however, how exactly to discharge this obligation towards my fellow members. What amount should I pay and to whom? This duty requires further specification if it is to be fulfilled in a meaningful way. Only a body that can calculate the correct amount and distribute it between all relevant members of the community, while possessing the ability to ensure its collection, can specify how to fulfil this duty, thanks to its de facto authority. Once the proper amount is set and the relevant directive is given to all relevant members of the respective community, obeying the directive will enhance my capacity to conform to reason and fulfil my obligation.

Indeed, coordination does play a role here, but this example is different from the previous ones, where the role played by collective action problems is intrinsic to the difficulty of conforming to reason. In this case, the problem is not collective action itself but rather the difficulty of discharging certain abstract prior obligations. The focus here is on the authority's ability to help us discharge obligations in a way we could not pursue without the directive. Collective action problems thus only play an instrumental role in the existence of a difficulty to discharge a prior obligation. Therefore, cases of further specification of obligations do not entail collective action problems by definition. But this example nevertheless describes a type of normative collective action problem. It is difficult to discharge the obligation precisely *because* we live in large, complex societies, and it is a collective action problem—in this case, funding a public good (as a result of a free rider situation, for instance)—that gave rise to the normative problem in the first place. Without collective action problems, this difficulty would not have been incurred, as it pertains to interacting with other individuals. Therefore, collective action problems, albeit instrumental in this context, are inherent to this type of difficulty of discharging obligations.

Analysis of the requirement for robust reason-giving and its application as described above leads to the observation that, when no prior obligation to obey exists, reasons for action can only be robustly given when normative collective action problems are involved. If there is no problem limiting our capacity for rational action, the command will fail to provide a reason for action in the robust sense—the sense relevant for authority. We would simply not be robustly given a reason to follow the command. The common element shared by all instances where the authority succeeds in giving a protected reason without relying on a prior obligation to obey is that they all have to do with how we behave and interact with other persons. One of the consequences of living in a complex society is a reduction in the ability to conform to reason in some situations, as the conduct

of others is often difficult to rely on. This reduction is manifested in collective action problems. Another implication of this analysis is that, when there is no prior obligation to obey, de facto authority is a prerequisite of legitimate authority.

Raz himself states that 'the case for the legitimacy of any political authority rests to a large extent on its ability to solve coordination problems and extricate the population from prisoner's-dilemma type situations.'[28] But, given the nature of robust reason-giving (more accurately, robust duty-giving of the sort authority engages in), it is difficult to see how in fact *any* authority, not just political ones, can be legitimate without this ability (again, prior obligations to obey put aside).

The main paradox of authority, which Raz called 'the moral problem' and Robert Paul Wolff referred to as 'the anarchist challenge',[29] addresses the incompatibility of authority and autonomy. Raz has made great advancements in settling these two concepts with the service conception. If we take autonomy seriously, we should aim to find a reasoning that can limit the relinquishment of self-judgment without forfeiting the benefits of authority. In my account, the justification for authority lies in the need for a device to help us overcome the obstacles stemming from the commonplace nature of interaction between individuals, ultimately diminishing our options for rational action, crippling the ability to achieve our own goals. Yielding to authority is the way—perhaps the only way, or best way—for humans to overcome these hurdles. Self-judgement will, in such cases, only limit our capacity for acting rationally, compared with being bound by directives. This means, for example, that on an island with only one person, no one can have authority.

This understanding is reinforced by the nature of preemption as an inherent feature of authority. The directive must normatively bind the subjects if it is to fulfil its role. It must guarantee the proper reaction of each member of the relevant community, and this is achieved by the binding power of the directive. In the absence of preemption, a reason for action might not be given at all. There may very well be theoretical authorities capable of producing reasons for belief, but these are not practical authorities that produce reasons for action. They can neither 'change things in the world'[30] nor produce duties by mere say-so. The nature of directives in normative collective action problems is that they must be treated as binding if they are to enable us to solve problems. It is this nature that brings the ability to morally bind subjects (or, at the least, create protected reasons).

A refined outline of the cases in which we are morally obligated to obey holds a theoretical advantage. Duties are a serious matter, and an account that sets a high bar for their production by mere say-so seems to balance the tension between

[28] ibid 21.

[29] Robert Paul Wolff, *In Defense of Anarchism* (New York: Harper & Row 1970). I understand Raz's moral problem of authority as reiterating what Wolff referred to as 'the anarchist challenge.'

[30] The Problem of Authority (n 2) 1034.

autonomy and authority in a satisfying manner.[31] This is consistent with our conception of liberty and seems to take on, with relative success, Wolff's challenge regarding the incompatibility of authority and autonomy. Only problems that could not be otherwise redressed in a significant way and which the redressing thereof ultimately serves the subjects can justify this (normatively speaking) drastic measure of relinquishing self-discretion that happens when one is subject to authority.

3. Objections

The above analysis of Raz's service conception might raise some objections. Relevant though they are to my interpretation, I will refrain from repeating all the general objections levelled against Raz's account and focus rather on difficulties that stand out in particular, in light of my suggestion to limit the service conception to normative collective action problems.

The first objection, raised by Scott Hershowitz,[32] is a general objection to the Razian account. According to this objection, the service conception leads to the odd conclusion that criminal law has no legitimate authority. This is due to the fact that most criminal laws simply set the floor for acceptable behaviour, while most of us are perfectly able to conform to our reasons and obligations to respect the lives and property of others without a directive ordering us to do so. In other words, the directive does not enable us to conform better to reason, leaving the NJT not satisfied. The implausible result of criminal law having no authority renders the NJT an unfitting criterion for authority.

However, asserting that criminal laws do not give reasons in the robust sense might be too hasty. Consider Hobbes and his State of Nature. In this State, rational agents acting by their rational capacity cannot achieve optimal results as they might if they could rely on the behaviour of others and secure valuable cooperation. Indeed, the reason to respect the lives of others might exist for them independently, but in a world where many others fail to act on this reason and threaten the agent (as the lack of the directive and its enforcement will have a causal

[31] A somewhat surprising advantage of this understanding lies in its appeal to consent-oriented theorists. There is no consent in this account, actual or hypothetical, and Raz famously denounces a consent-based theory of authority. But hypothetical consent theorists focusing on normative notions of consent may find this appealing. See David Estlund, *Democratic Authority* (Cambridge: Cambridge University Press 2007). Situations like free rider problems and coordination problems are cases in which the subject's actual consent isn't given not because she finds it against her interests or reasons for action, but because she does not have the capacity to act on them. Given the fact that a lack of obedience can manifest itself in results harmful to society, normative consent can be said to be given in these cases.

[32] See Accountability and Political Authority (n 5).

effect in the world), the agent's reason for self-preservation is likely to outweigh the contradicting reasons.

As we know from the familiar game theory analysis of *Leviathan*, outside interference is needed to solve this prisoner's dilemma successfully and avoid living a 'nasty, brutish, short'[33] life. This is a collective action problem in the narrow economic sense, focusing on the interest of the agent, but self-interest is a normative reason for action as well, and therefore it is also included in the wider sense that I address. In this instance, regardless of the reason we all have to respect the lives of others, the directive increases the capacity to conform to reason by eliminating risks posed from interacting with others. The issuing of the directive, along with a de facto ability for enforcement, is what coordinates and gives the agent the ability to rely on the behaviour of others. This makes possible developing a life free of constant fear and benefiting from cooperation with others. This makes for better conformance to reason, thus constituting reason-giving in the robust sense. Consequently, the typical prohibitions on murder, robbery, assault, etc. (accompanied by the enforcement of which) are authoritative by virtue of their coordinative quality.[34]

Alternatively, one may remain unconvinced by the Hobbesian example. But perhaps the fact that many criminal ordinances do not have authority over their subjects is not such an implausible conclusion after all. Most persons do not need the law prohibiting murder to enable them better to conform to reasons applying to them anyway (namely, an obligation to respect the lives of others). If one's reason to respect another individual's life is indeed sufficient for her to conform to reason and not take the life of another, her actions are the same as they might be, were she to follow the directive. If the end result is the same, the lack of authority of criminal laws has no physical impact in the world. Albeit applying only to causal rather than normative aspects, this already makes the conclusion of criminal law having no authority less scandalous than it may have sounded at first. Similarly, even if criminal law has no authority over the general public (or at least the law-abiding majority thereof),[35] it does not follow that the lack of authority of these provisions leads to a situation in which violating them is somehow permissible or

[33] Thomas Hobbes, *Leviathan* (first published 1651, Indianapolis: Hackett Publishers, 1992).

[34] It is possible that not all criminal prohibitions always carry such coordinative quality. Perhaps legal norms such as prohibiting certain drug use or requiring wearing a seat belt in a car generally lack this quality. It seems quite plausible to me that a legal system would have authority in some cases but not in others.

[35] It can be argued that criminal law has authority over certain persons in their official capacity, such as judges, prosecutors, or other law enforcement personnel. This however is not the critique Hershovitz aims for, as I understand it. His critique refers to the common notion that criminal prohibitions such as the prohibition on murder or theft must be obeyed by the general public.

less morally grave. Nor does it follow that the use of coercive force by the state on those who commit such acts is not justified.[36]

Another part of Hershovitz's argument includes a person with a duty to financially support her family who goes to hear the advice of a financial expert as to how to invest money for the future in the best way possible. Abiding by the expert's utterance as to which fund to invest in would make an investor conform to reason better, thus creating a binding directive from a financial advisor—an implausible result.[37] In addition, is it not arbitrary or even random that I am obliged to obey this particular expert who happens to advise me? Isn't it odd that the 'right to rule, understood as correlated with the obligation to obey on the part of those subject to the authority',[38] can be so arbitrarily exercised?

Not necessarily. Firstly, again following Enoch, I suggest that the right to rule needs to be understood simply as the capacity to create an obligation to follow its directives. It does not contain an element of directionality. This, as Enoch contends, is because the duty to obey is not owed to authority itself. To demonstrate this, Enoch describes an arbitrator,[39] agreed upon by two opposing parties and possessing practical, legitimate authority. The arbitrator has produced a decision in favour of one of the parties, but the losing party refuses to pay. To whom does the losing party owe the duty to obey? Who is the injured party whose rights have been breached? It is much more plausible to say that the injured party here is not the arbitrator herself. If anyone, it is the winning party, even though she is not the possessor of authority in this case. Consequently, the right to rule should not be understood as obeying authority *as an entity*. As Raz puts it, one cannot *be* a practical authority; one can only *have* practical authority.[40]

Without directionality, the randomness of attaining the right to rule is less of a surprise, since no special standing with the authority is necessary. It may come as no great surprise that the right to rule is achieved with what can be viewed as random or arbitrary factors. It might be the case that the financial advisor was simply the first option to come up in the telephone directory. But this does not seem to hold particular relevance for the authoritativeness of the directive. It is unlikely that the way some people have gained political authority is excused of elements

[36] It might be worthwhile to stress that I am not positively claiming that criminal law does not require any justification. Rather, I raise a more modest negative point: that Hershovitz's objection is less powerful than it might seem at first glance, in the sense that even if criminal law does not generate duties according to the service conception, this would not lead to counter-intuitive conclusions such as murder or theft being morally permissible, or that imposing criminal sanctions is not justified (see, also, Joseph Raz, 'The Obligation to Obey: Revision and Tradition' (1985) 1 Notre Dame JL Ethics & Pub Pol'y 139, 143–144). However, other counter-intuitive ramifications of this assertion might remain, and are not tackled in this article.

[37] The Role of Authority (n 4), 10.

[38] Joseph Raz, *The Morality of Freedom* (1986) 23.

[39] See Enoch (n 22) 28–29.

[40] The Problem of Authority (n 2) 1032–1037. The same is not true of theoretical authorities. With them, someone *is* the authority. See ibid.

such as chance or opportunity. If the main role of legitimate authority ultimately is to serve its subjects, focusing on elements other than this feature as a necessary condition for authority seems ad-hoc, in this aspect.[41]

This understanding contradicts Darwall's and Hershovitz's approach, which assumes the existence of some normative standing or status between ruler and ruled. For Darwall, the authoritativeness of the command is expressed in the accountability of the subject towards the authority.[42] Similarly, Hershovitz posits that the 'ordinary way' of understanding the right to rule is in the form of the particular relationship between the ruler and the ruled.[43] For them, authority necessarily includes an element of directionality: the duty to obey is owed *to the authority*. Hershovitz and Darwall alike share this conception. What can be the source of this anthropomorphic view of authority?

Perhaps because directives usually come from a person or perhaps because of the ambiguity of natural language, authority can be misconstrued to have person-like attributes. Indeed, authority can be tightly connected to the identity of the person giving directives—so much so that they may sometimes seem interrelated. This holds particularly true if authority stems from a prior obligation to obey a certain person, perhaps a parent. But these are usually those cases where a prior obligation to obey exists and the NJT simply 'confirms' authority (the 'boring' type of cases discussed earlier, when the NJT is not an interesting criterion for determining an obligation to obey). Here too, the existence of practical authority is more of an inference drawn from the special relationship and the obligation it already contains, rather than an inherent part of the relationship.

Authority is an intangible instrument at our disposal (admittedly, an instrument that must be used by people if it is to fulfil its purpose), and it is people who are ultimately the possessors of practical authority. Nevertheless, authority is an instrument. The person issuing the directive is not an authority; he only *has* authority. Indeed, Raz convincingly states that he is more used to the idea that the authority is accountable to its subjects than the other way around.[44] This does not contradict the fact that authority can stem from prior obligations rooted in special relationships.[45] As stated, standings involving prior obligations can be sufficient to establish authority, but not every instance of authority necessarily entails it. This

[41] Raz leaves room for such considerations, as an alternate sufficient criterion. He explains that along with the Normal Justification Thesis, which is the normal way to accept authority, there are other, 'deviant' reasons for one to accept authority. This is beyond the scope of the NJT and this Article. See Authority and Justification (n 11) 19–21.

[42] See Darwall (n 3).

[43] Hershovitz has a slightly more nuanced account than Darwall's, based upon the roles individuals play in practice (see The Role of Authority (n 4) 11–19). The core point, however, remains the same, since the roles are 'played' by people.

[44] On Respect (n 9) 299.

[45] Joseph Raz, 'Promises and Obligations', in P.M.S. Hacker and Joseph Raz (eds), *Law, Morality and Society: Essays in Honor of H.L.A. Hart* (Oxford: Clarendon Press, 1977) 228.

narrow understanding coincides with the instrumental role of authority, as well as with Raz's wording.[46] It is consistent with the view of authority as a minimalistic mechanism, a device with no personality or human characteristics, intruding as little as possible, only when people have problems that could not otherwise be redressed in a significant way.

Secondly, and more importantly, it may be the case that the financial advisor does not have the capacity to impose duties according to the service conception at all. The expert did not in fact succeed in creating a reason for action, as the utterance itself only instructed the investor on how best to invest. The investor already had the rational capacity to choose to invest in that particular fund; she simply hadn't had enough information. The utterance itself is not what made it possible for her to conform to reason and invest in this particular fund. But how is this different from the difficulty to discharge obligations because we do not know how to do so, as in the tax example discussed previously? The financial advisor scenario is different because, in the tax example, the authority's setting of the correct amount and system of collecting, enforcing and coordinating could not have been achieved without the authority, as the correct amount for an individual to pay depends, among other things, on what his peers would pay. In this sense, the very existence of a 'proper' amount to pay is only possible thanks to the authority and its issuing of the directive. By contrast, the financial advisor has merely given the investor a reason to hold the *belief* that investing in a particular fund will make her best conform to reason; i.e., a reason for belief, rather than a reason for action. Accordingly, this is a case of theoretical, rather than practical, authority.[47]

Another objection focuses on a further implausible implication: the illegitimacy of parental authority. Arguably, parents have—in general—authority over their children. But if the above analysis regarding the conditions in which reasons are robustly given is correct, it is difficult to see how parents give a reason for action in the robust sense when they issue directives. Parents, at least those of a single child, do not coordinate anything, nor does it seem that they specify a prior unclear obligation of some sort. The NJT seems like a good account of political authority, but when it comes to parental authority, the NJT (or at least my account of it) apparently falls short. There does not seem to be any normative collective action problem that parental authority aims to solve. Similarly, the NJT seems to fall short in explaining why specific parents have authority over specific children.

Indeed, the collective action problems previously discussed seem beside the point here, and yet it appears very plausible that parents have authority over their children. (At least some parents, some of the time). But we have also said that this is not the only way to meet the NJT criterion and establish authority. There is another way to do so. As you may recall, there are instances where the NJT does not perform any interesting theoretical work. Those are the cases where

[46] Joseph Raz, *The Authority of Law* (Oxford: Clarendon Press, 1979) 16–20.
[47] On Respect (n 9) 301.

there is a preexisting, independent obligation to obey someone or something. This preexisting obligation constitutes a reason for action, a reason that would unsurprisingly be best conformed to by obeying the directive. But in this 'boring' scenario, we must still safely establish that there is a prior obligation for a daughter to obey her parents. Can this be achieved? The answer to this question requires a discussion that is beyond the scope of this article, but I will argue in short in the following paragraph that the answer is yes. It should be noted, however, that if an independent, prior obligation of a daughter to obey her parents cannot be determined, then it must be conceded that, according to my account and contrary to common intuition, parental authority is false, with all the bullet-biting rewards of this conclusion.

In his example, Hershovitz uses the example of a stranger with excellent parenting skills approaching another parent's child and ordering them around. He uses it to establish that better parenting skills, i.e. making the child conform better to reason, are not enough to establish authority (thus disproving the NJT); rather, it is the special moral relationship between a parent and their children that establishes it.[48] Hershovitz agrees that parents and children have a special standing between them, one that can give rise to duties. Now, once we remember the 'boring' scenario and assume a prior obligation of children to obey their own parents, this example poses no special problem for the NJT.

4. Conclusion

The notion that the main role of authority is to solve collective action problems is not a new one. However, Raz avoids any explicit reference to collective action problems as the raison d'être of authority. My aim was to show that when there is no special, prior obligation to obey, the solving of such problems is the only path for establishing legitimate authority, according to Raz's own account. Another observation to be drawn from the analysis presented here is that, in these types of cases, legitimate authority can only exist when there is de facto authority, since only de facto authority has the physical capability to coordinate and solve these problems. This understanding of authority has the advantage of capturing the significance of autonomy while still fitting within our intuitive grasp of the concept of authority. Raz's account seems to support this understanding, or at least not to contradict it.

[48] See Accountability and Political Authority (n 5) 13.

Limiting the Use of Cautions: Avoiding 'Cautions Culture' and Collateral Consequences

Carlene Miller[1]

1. Introduction

In the years 2013 and 2014, nearly 400,000 cautions were issued by police across England and Wales, including for serious offences involving children, sexual acts, and weapons.[2] This article argues that limiting the use of cautions supports the goals of both the Government and other proponents for 'tough on crime' policies, as well as would-be defendants and other advocates for pro-defendant policies. In October 2013, then-Minister of Justice Chris Grayling announced plans to reform what he and others believe is a 'cautions culture': the over-cautioning of serious and often repeated offences, resulting in what is perceived as nothing more than 'a slap on the wrist' for offenders.[3] The reforms culminated in the Criminal Justice and Courts Act 2015, which partially limits police's ability to caution for certain offences.[4] Meanwhile, some defence lawyers

[1] J.D. candidate, Notre Dame Law School. I would like to thank the editors and wish them luck for the future of this journal. This article was written during my year at the Notre Dame Law in London Programme, from which I have now returned to the United States. I would also like to thank Professor Penny Darbyshire of Kingston Law School and Notre Dame Law in London for providing excellent topic and research advice and Lauren Jennings and Kiri Abadir of JD Spicer Zeb Solicitors for teaching me what cautions are in the first place. Lastly, I would like to thank Trevor Stevens for his help with this article.

[2] Stefano Ruis, 'The Hidden Mischief of Police Cautions' (*The Justice Gap*, 5 September 2014) <http://thejusticegap.com/2014/09/hidden-mischief-police-cautions/>; Brooke Perriam, 'Grayling Promises to End "Cautions Culture"'(*The Justice Gap*, 4 November 2014) <http://thejusticegap.com/2014/11/grayling-ends-soft-option-cautions/>.

[3] Perriam (n 2); Tom Wright, 'Chris Grayling Announces Changes to Police Cautions' (*The Justice Gap*, 2 October 2013) <http://thejusticegap.com/2013/10/chris-grayling-announces-changes-police-cautions/>.

[4] Criminal Justice and Courts Act 2015, ss. 15–17.

and other defence proponents believe the current caution system is also in need of reform as would-be defendants may not fully understand that a caution carries with it a criminal conviction and various collateral consequences that may affect future employment, character evidence given in future court proceedings, and other aspects of life.

In perhaps a rare occurrence, the goals of advocates both for and against 'tough on crime' policies can be met by limiting the use of cautions. In limiting the use of cautions, the Government can lessen the prevalence of the 'cautions culture' that 'tough on crime' proponents believe encourages recidivism and further offending. Simultaneously, would-be defendants will receive fewer unadjudicated convictions on their criminal histories, thus avoiding the collateral consequences of those convictions in the areas of employment background checks and bad character evidence used against them in future legal matters. Limiting the use of cautions thus furthers the Government's goal to disincentivise further offending while concurrently avoiding the over-penalisation of would-be defendants brought on by the collateral consequences of cautions.

2. Development of the Use of Cautioning

The practice of cautioning began with juveniles in an effort to limit juveniles' exposure to the criminal justice system.[5] The Children and Young Persons Act 1969 grants the original statutory authority to caution.[6] By 1981, the Parliamentary All-Party Penal Affairs Group supported the view that youth cautions were an excellent method to address delinquency if the youth is not a persistent delinquent, a police warning in a formal setting would be sufficiently impactful, the family had been alerted, and the youth could be connected with an agency that could assist in alleviating the factors making the youth delinquent.[7] A 1983 study showed that youth cautions appeared to be achieving their intended result, since juveniles receiving cautions were less likely to re-offend than those who were prosecuted.[8] Other justifications for cautioning are that it avoids stigmatising juveniles, connects the juvenile's family with social services, and saves police and court resources from being squandered on trivial offences.[9]

Juvenile cautions were, however, granted inconsistently, varying in the type of offences cautioned and the number of cautions given to a single offender.[10] In response to these inconsistencies, the Home Office issued guidelines in 1985 that cautions were to be given only when the seriousness of the offence falls short of the need for prosecution and where there was: '(a) sufficient evidence to prove the

[5] Ronald Bartie, 'A Deviation from Crime' [1990] 140 NLJ 1494.
[6] ibid.
[7] ibid.
[8] ibid.
[9] Sean Enright, 'Charge or Caution?' [1993] 143 NLJ 446.
[10] ibid.

case, (b) the juvenile admitted the offence, and (c) the juvenile's parents had given their agreement to this course of action.'[11] The Crime and Disorder Act 1998 further required that for a caution to be given, the police must determine that the child has committed an offence, that there is a realistic prospect of conviction, that the offence has been admitted, that the child has no previous convictions, and that it is not in the public interest to prosecute.[12] Neither the consent of the child nor the child's appropriate adult was a condition.[13] If the child had been previously cautioned in the last two years, a caution could not be given.[14] A two-category system of youth and youth-conditional cautions was then created by The Legal Aid, Sentencing, and Punishment of Offenders Act 2012.[15] Cautions would now be given where the child admits to the offence, the police decide there is sufficient evidence to charge and that the child should not be prosecuted, the caution is given in the presence of an adult, and the implications of the caution are explained.[16] However, consent was not required nor were there safeguards preventing the adult from urging the child to confess.[17] These changes were introduced in response to the case of *R v Durham Constabulary*, where a 14-year-old boy was cautioned for sexual assault, but was not told until a week later about his obligation to register on the Sex Offenders Registry.[18] The House of Lords quashed the appeal, though Baroness Hale did criticise the lack of a consent requirement.[19]

Adult cautions were considered only a 'possible course of action' in the 1985 Home Office guidelines and were only considered suitable for elderly or vulnerable adults.[20] The Home Office's stance drastically changed by 1990, when their circular announced that adults should not 'be excluded from cautioning by reason only of their age.'[21] Adult cautions were to be given only when there was 'an admission of guilt' and 'sufficient evidence to prove the charge.'[22] The guidelines iterated that in assessing whether a caution should be given, the nature of the offence, the likely penalty resulting from prosecution, and the offender's age, health, attitude, and previous record should all be considered.[23] By the early 1990s, adult cautions were

[11] ibid; Home Office Circular 14/1985.

[12] Crime and Disorder Act 1998, s. 65(1); Jon Robins, 'Youth Cautions and the Slap on the Wrist' (*The Justice Gap*, 16 November 2012) <http://thejusticegap.com/2012/11/youth-cautions-and-the-slap-on-the-wrist/>.

[13] Robins (n 12).

[14] ibid; Crime and Disorder Act, s. 65(2).

[15] Legal Aid, Sentencing and Punishment of Offenders Act 2012, ss. 135–138 ['2012 Act']; Ronald Ellis & Stuart Biggs, 'Simple Cautions' [2013] 5 AR 6, 7.

[16] 2012 Act, s. 66ZA(1)–(4); Ellis and Stuart (n 15).

[17] Ellis and Biggs (n 15).

[18] *R v Durham Constabulary* [2005] UK 21.

[19] ibid [39].

[20] Home Office Circular 14/1985; Ellis and Biggs (n 15).

[21] Home Office Circular 59/1990; Ellis and Biggs (n 15).

[22] ibid.

[23] ibid.

widely being used and the range of offences quickly expanded to include theft, shoplifting, public order offences, minor assaults, criminal damage, and possession of controlled drugs.[24] While cautions for drug possession were originally limited to cannabis, a 1993 Metropolitan police directive expanded cautions to include class-A drugs like cocaine and heroin.[25] By 2012, adult cautions were regularly given for child prostitution and pornography, cruelty to or neglect of children, and other indictable-only offences.[26]

The use of adult cautions was further expanded by the Criminal Justice Act 2003, which created simple and conditional cautions,[27] while providing the Crown Prosecution Service with greater discretion in cautioning.[28] A conditional caution is given when the offender has made an admission and, having the effects explained to him by an authorised officer, agreed to the caution,[29] which carries conditions such as compensation, drug addiction programmes, apologies, or attendance at victim counseling programmes.[30] Conditional cautions have also been tested specifically for female sex workers, with the intention to divert women from prison and towards women's centres, which provide advice and educational courses.[31] More serious offences continued to be prescribed as simple, rather than conditional, cautions, and a 2008 Home Office circular reemphasised that simple cautions are to be used for only low-level offences.[32]

These concerns prompted a review of cautioning by Justice Secretary Jack Straw in December 2009.[33] In October 2013, Minister of Justice Chris Grayling announced that cautions for all indictable-only offences would be banned and

[24] Enright (n 9) 446.

[25] ibid.

[26] Catherine Baksi, 'Grayling Pledges No More "Slaps on Wrist" for Rapists and Child Sex Offenders' (*Law Society Gazette* , 1 October 2013) <http://www.lawgazette.co.uk/practice/grayling-pledges-overhaul-of-cautions/5037948.article>.

[27] Criminal Justice Act 2003, ss. 22–23.

[28] Criminal Justice Act 2003, s. 23B, as inserted by the Criminal Justice and Immigration Act 2008; Ken Macdonald, 'The New Code for Crown Prosecutors' [2005] 155 NLJ 12.

[29] Criminal Justice Act 2003, ss. 22(1), 23.

[30] 'Crime Brief' (*New Law Journal*, 2 August 2007) <http://www.newlawjournal.co.uk/nlj/content/crime-brief-17>; Under s. 22(3) of the Criminal Justice Act 2003, the conditions which may be attached to conditional cautions are those which have one or more of the following objects: (a) facilitating the rehabilitation of the offender; (b) ensuring that the offender makes reparation for the offence, and; (c) punishing the offender.

[31] 'Conditional Cautions Will Keep Women Out of Prison' (*New Law Journal*, 10 July 2008) <http://www.newlawjournal.co.uk/nlj/content/conditional-cautions-will-keep-women-out-prison>.

[32] Home Office Circular 16/2008; 'Criminal Litigation' (*New Law Journal*, 25 July 2008) <http://www.newlawjournal.co.uk/nlj/content/law-digest-192>.

[33] Catherine Baksi, 'Government to Review Use of Cautions' (*Law Society Gazette*, 14 December 2009) <http://www.lawgazette.co.uk/news/government-to-review-use-of-cautions/53529.article>.

offenders may face consequences, such as fines.[34] The governmental guidance of Grayling and his successor Michael Gove was made statutory by the Criminal Justice and Courts Act 2015, which mandated that cautions should not be given, save for in exceptional circumstances, if the offence is indictable-only[35] or a specific either-way offence[36] or if the offender has been cautioned in the past two years.[37] The indictable either-way offences specified include offences related to crimes against children, crimes involving weapons, sexual crimes, and class-A drug crimes.[38]

3. Effects of Cautioning on Would-Be Defendants

While these various reforms over the years have grown out of a concern for how the rampant use of cautions harms victims and perceived law-abiding citizens, reforms have rarely addressed the effect of cautions on the would-be defendants who receive them. Cautions result in a number of collateral consequences, which would-be defendants often do not fully comprehend, especially if they are juveniles.[39] Many cautions are accepted hastily without any legal representation, due to the LASPO 2012 cuts to legal aid and would-be defendants' eagerness to leave the police station.[40] Cautions can prevent travel abroad, especially to countries with strict immigration policies like the United States.[41] Cautions are also likely to affect sentencing in future prosecutions, though the research is inconclusive because Ministry of Justice sentencing data does not separate prior convictions from prior cautions.[42] The most serious collateral consequences of cautions are their effects on would-be defendants' employment criminal background checks and bad character evidence in future legal matters. These collateral consequences will be discussed in turn.

[34] ibid; Perriam (n 2). However, in a statement following Grayling's announcement, Surrey Police Chief Constable Lynne Owens made sure to clarify that 'the use of simple cautions for indictable-only offences represent a fraction of 1% of the total issued. Therefore, the police service would take the view that these are only used in exceptional circumstances currently.'

[35] Criminal Justice and Courts Act 2015, s. 17(2) ['2015 Act'].

[36] 2015 Act, s. 17(3).

[37] 2015 Act, s. 17(4)(b).

[38] Anthony Edwards, 'Criminal Law Changes' (*Law Society Gazette*, 15 June 2015) <http://www.lawgazette.co.uk/law/legal-updates/criminal-law-changes/5049333.article>; Ministry of Justice Guidance, 'Simple Cautions' (13 April 2013).

[39] Ellis & Biggs (n 15); Ruis (n 2).

[40] Ellis & Biggs (n 15) 9.

[41] Julian V. Roberts & Jose Pina-Sanchez, 'Previous Convictions at Sentencing: Exploring Empirical Trends in the Crown Court' [2014] 8 CLR 575, 582; David Sleight, 'Treat Reforms with Caution' (*Law Society Gazette*, 5 November 2014) <http://www.lawgazette.co.uk/law/practice-points/treat-reforms-with-caution/5044849.article>.

[42] Roberts & Pina-Sanchez (n 41) 582.

4. Cautions and Background Checks

Accepting a caution mars a would-be defendant's criminal record with a conviction, which can be discovered during a criminal background check and create grave consequences for the would-be defendant's employment. The argument for including cautions in criminal record checks is that in accepting a caution, an individual admits their guilt and the caution should be treated as if it were a conviction, without any question of evidence having been 'inconclusive.'[43] However, not all share this view; in a 2009 lecture, Lord Justice Leveson, President of the Queen's Bench Division, posited that:

> In issuing an out of court disposal, the police are essentially acting as prosecutor and judge, outside the environment of an open court. Although these disposals are not convictions, they are kept on record and, at the least serious end, can risk criminalizing people who on a one-off occasion do something out of character, and who feel the quickest thing to do is accept the penalty or caution that is being proposed by the police, even if further analysis might have revealed no offence.[44]

Criminal record checks are required for all work that involves children or vulnerable adults—even unpaid, voluntary work, such as scout leading—and can include work in healthcare, law, and the Civil Service.[45] Doctors, lawyers, registered financial practitioners, and armed forces personnel who are cautioned may face separate investigation and disciplinary hearings.[46]

In attempts to lessen the severity of these consequences, there have been a number of reforms made to the criminal record check system. Criminal record checks are currently conducted by the Disclosure and Barring Service, which was launched in December 2012 as a merger of the Criminal Records Bureau and the Independent Safeguarding Authority.[47] Previously, a criminal record check revealed current and spent convictions (including cautions), reprimands, and warnings.[48] Originally, under the Rehabilitation of Offenders Act 1974, cautions

[43] Richard Scorer, 'Blacklisted' [2006] 156 NLJ 125.

[44] Rachel Rothwell, 'Out of Court Disposals Warning' (*Law Society Gazette*, 9 December 2010) <http://www.lawgazette.co.uk/news/out-of-court-disposals-warning/58440.article>.

[45] Helen Hart, 'Checking Up: Are Criminal Records Bureau Checks Too Onerous? Ask Helen Hart' (*New Law Journal*, 15 February 2008) <http://www.newlawjournal.co.uk/nlj/content/checking>; Ellis & Biggs (n 15) 7.

[46] Ruis (n 2).

[47] Ellis & Biggs (n 15) 7.

[48] Hart (n 45).

became spent after the offender did not recidivate after a specified period of time.[49] This policy was changed in the Criminal Justice and Immigration Act 2008, which spent simple cautions as soon as they were imposed and spent conditional cautions three months after their imposition if conditions were met.[50] A filtering scheme introduced in May 2013 no longer discloses minor convictions and cautions after six years for adults and after two years for juveniles.[51] The filtering scheme has, however, a number of exclusions for cautions related to listed offences and cautions issued to would-be defendants with previous convictions.[52] Many people with multiple minor cautions will continue to have cautions disclosed for the rest of their lives.[53] It should be noted that obtaining multiple minor cautions, so as to be excluded from the filtering scheme, can derive from something as simple as being overpaid benefits for two consecutive months and receiving one caution for each month.[54]

The courts have also weighed in on the employment consequences of cautions appearing on criminal records. In 2005, the Information Tribunal ruled in *Chief Constable of West Yorkshire, South Yorkshire and North Wales v Information Commissioner* that old records could be retained for 'policing purposes and the administration of justice' but were not to be disclosed for other purposes, such as vetting.[55] Practically, this did not occur until statutory intervention by the Crimes and Courts Act 2013, though some cautions, for serious sexual and violent offences, will always be disclosed.[56]

Disclosure of cautions in criminal record checks also raises issues concerning Article 8 of the European Convention on Human Rights, as the European Court of Human Rights ruled in *M.M. v The United Kingdom* that cautions are a part of a person's private life.[57] This case arose out of a Northern Irish caution, which has some procedural differences to English cautions but, nonetheless, raises Article 8

[49] Rehabilitation of Offenders Act 1974, Schedule 2, s. 3.

[50] Criminal Justice and Immigration Act 2008, s. 49 & Schedule 10; Anthony Edwards, 'Criminal Law Roundup: More than Just the Usual Suspects' (*Law Society Gazette*, 22 October 2009) <http://www.lawgazette.co.uk/law/criminal-law-roundup-more-than-just-the-usual-suspects/52791.article>.

[51] Jamie Grace, 'Old Convictions Never Die, They Just Fade Away: The Permanency of Convictions and Cautions for Criminal Offences in the UK' [2014] 78(2) JCL 121, 131.

[52] Ruis (n 2).

[53] Christopher Stacey, 'Filtering of Cautions and Convictions Doesn't Go Far Enough' (*The Justice Gap*, 21 June 2014) <http://thejusticegap.com/2014/06/filtering-cautions-convictions-doesnt-go-far-enough>; 'No Place for Cautions' (*New Law Journal*, 19 June 2014) <http://www.newlawjournal.co.uk/nlj/content/no-place-cautions>.

[54] Stacey (n 53).

[55] *Chief Constables of West Yorkshire, South Yorkshire and North Wales Police v Information Commissioner* [2005] UKIT DA 05 0010 (12 October 2005), para. 220.

[56] Anthony Edwards, 'Legislation and Case Law' (*Law Society Gazette*, 30 September 2013) <http://www.lawgazette.co.uk/law/legal-updates/legislation-and-case-law/5037881.article>.

[57] *M.M. v United Kingdom* (App. no. 24029/07) [2012] ECHR 1906.

issues.[58] English courts examined how disclosure engaged Article 8 in *R. (T) v Chief Constable of Greater Manchester & Others*.[59] In 2002, T, at the age of 11, admitted to the theft of two bicycles and was given two cautions; in 2010, T applied for a sports studies course, which involved contact with children, thus requiring a criminal background check that revealed his two cautions.[60] The UK Supreme Court upheld the Court of Appeal ruling that the indiscriminate statutory regime requiring the disclosure of all cautions violated Article 8 on two grounds: that a caution takes place in private, making the caution protected personal information, and that the impairment of employment opportunities affects a person's ability to enjoy private life.[61] The Supreme Court reasoned that the lifelong disclosure of minor cautions was 'disproportionate', 'not necessary in a democratic society', and 'not based on any rational assessment of risk.'[62] Further support for disclosure reform was seen the day after the Supreme Court's ruling, when a Parliamentarian's Inquiry led by Lord Carlile QC published a wide range of recommendations, including ways that juvenile criminal records should be dealt with.[63] Specifically, the inquiry recommended that filtering rules should be extended to offences that resulted in a prison sentence of six months or less and that child offenders should receive lifelong anonymity.[64]

5. Bad Character Evidence

Cautions also result in unforeseen consequences in the arena of bad character evidence in future legal matters, most often in would-be defendants' future criminal trials. The Court of Appeal has ruled that cautions can be used as evidence of bad character because acceptance of a caution requires an admission of guilt.[65] District

[58] ibid [159]–[174].

[59] *R(T) v Chief Constable of Greater Manchester & Others* [2013] EWCA Civ 25; Adam Jackson, 'Case Comment: Criminal Records, Enhanced Criminal Records Certificates and Disclosure of Spent Convictions: Impact of ECHR, Article 8' [2014] 78(6) JCL 463.

[60] ibid; 'Criminal Records—Police Act 1997 ss. 113A and 113B—European Convention on Human Rights Art. 8—Compatibility—Rehabilitation of Offenders Act 1974 (Exceptions) Order 1975—Whether *Ultra Vires*' 7 AR 1; Sam Thomas, 'Case Comment: The Supreme Court Judgment in R. (on the Application of T) v Chief Constable of Greater Manchester and the Effect on Professional Regulators' [2015] 2 CLR 149.

[61] *R(T) v Chief Constable of Greater Manchester Police & Others* [2014] UKSC 35; [2014] WLR (D) 271; Catherine Baksi, 'Disclosure of Cautions Breaches Privacy Rights, Supreme Court Rules' (*Law Society Gazette*, 18 June 2014) <http://www.lawgazette.co.uk/law/disclosure-of-cautions-breaches-privacy-rights-supreme-court-rules/5041734.article>; Ellis & Biggs (n 15) 8.

[62] Stacey (n 53).

[63] Christopher Stacey, 'Filtering of Cautions and Convictions Doesn't Go Far Enough' (*The Justice Gap*, 21 June 2014) <http://thejusticegap.com/2014/06/filtering-cautions-convictions-doesnt-go-far-enough>; 'No Place for Cautions' (*New Law Journal*, 19 June 2014).

[64] Alan Travis, 'Children with criminal past should be given clean slate at 18, says MPs' (*The Guardian*, 19 June 2014) <http://www.theguardian.com/law/2014/jun/19/children-criminal-past-clean-slate-18-say-mps>.

[65] J. R. Spencer, 'Evidence of Bad Character—Where We Are Today' [2014] 5 AR 5.

Judge Gareth Branston correctly criticises this reliance on a would-be defendant's admission of guilt as justification for using cautions as bad character evidence.[66] Branston has been very critical of the use of cautions as bad character evidence, principally relying on the Criminal Justice and Immigration Act 2008's amendments to the Rehabilitation of Offenders Act 1974.[67] The amendments provide that a person given a caution be treated as if they had not committed an offence once the caution is spent, and that no evidence shall be admissible to prove a caution had been given; this seems to be limited to civil matters, however, as there exists an exception for admission of such evidence in criminal proceedings.[68] Branston argues that cautions are not misconduct, but merely evidence of misconduct, which is hearsay in criminal proceedings within the Criminal Justice Act 2003.[69] There is a counter-argument, particularly espoused by Professor J. R. Spencer, that cautions constitute an admission exception to hearsay because embedded in a caution is the fact that the defendant has confessed to the offence.[70] Branston rightly rejects this argument on the grounds that caution admissions strain the meaning of confession in criminal proceedings, that describing a previous caution as a confession is unsupported by authority, and that a confession is only admissible if made by a defendant and deployed by the prosecution or a co-defendant.[71]

6. Conclusion

Though it might seem paradoxical, limiting the use of cautions via Government reforms might be in the best interests of both the Government and would-be defendants, though not without implications for both. A conceptual trade-off exists between benefits the Government and would-be defendants receive and the resulting implications.

The Government benefits from limiting the use of cautions by furthering its goal of deterring future offences through what it deems to be adequate sentencing and punishment. Almost since the inception of juvenile cautions in 1969, there has been a steady Government effort to reform and limit their use. Despite Government efforts to reign in their use, cautions have continuously increased and expanded in number. The Government and other supporters of 'tough on crime' policies have viewed this continual increase in cautions as evidence that cautions are not an effective deterrent; in their eyes, too many would-be defendants evade adequate punishment and thus continue to offend. Limiting the use of cautions furthers the Government's goal of adequately punishing offenders in order to deter future offences.

[66] Gareth Branston, 'A Reprehensible Use of Cautions as Bad Character Evidence?' [2015] 8 CLR 594, 596.
[67] ibid.
[68] ibid.
[69] ibid.
[70] J. R. Spencer, 'Cautions as Character Evidence: A Reply to Judge Branston' [2015] 8 CLR 611.
[71] Branston (n 66) 600–01.

This has implications for the Government, however, as fewer cautions means that the Government must invest the time, money, and resources into prosecuting more offences. The extensive use of cautions has provided an inexpensive alternative to adjudication while still punishing offenders. By limiting the use of cautions, the Government will either need to allot more funds and resources to the Crown Prosecution Service ('CPS') to prosecute these offenders or accept that the CPS will be further constrained in their prosecution decisions. Practically, more offenders may evade punishment if cautions are limited and prosecutorial discretion is constrained by further resource limitations.

As has been discussed at length in this Article, the increased use of cautions over the years has amplified the collateral consequences of cautions, creating grave impacts on employment opportunities, future legal cases, and other matters of life for would-be defendants, who are often under-informed of these consequences. Limiting the use of cautions would protect would-be defendants from these collateral consequences. However, this also has implications for would-be defendants in that it increases their likelihood of becoming actual defendants facing actual prosecution. A defendant who may have only received a caution resulting in collateral consequences before may now face prosecution and an actual sentence. However, seeing as how all would-be defendants are already experiencing some punishment via collateral consequences, more prosecutions would at least result in some acquittals and would allow some defendants to avoid punishment altogether. Removing significant cautioning power from the police and CPS might also force their hands to use their discretion to prosecute only the most worthwhile offences.

In this way, limiting the use of cautions will advance the goals of both the Government and pro-defendant advocates opposed to 'tough on crime' policies and better protect the rights and interests of the very people who face prosecution by the Government.

Prayer for Relief: Saguenay and State Neutrality toward Religion in Canada

Ravi Amarnath[1]

Brian Bird[2]

1. Introduction

TO WHAT EXTENT can the state, in carrying out its functions, profess or favour one religious tradition over another? This question was at the heart of the decision of the Supreme Court of Canada in *Mouvement laïque québécois v Saguenay (City)*.[3] The Court decided that the Canadian state bears a duty of neutrality in matters of religion, which means it cannot profess or favour one religious tradition over another. This Article discusses the consequences of how the Court articulated the duty of neutrality in Canada, and, in particular, how it pertains to deriving a meaning for a 'secular state'.

2. Freedom of Religion in Canadian Constitutional Law

To begin, we will briefly summarise how freedom of religion has developed in Canadian constitutional law since the advent of the *Canadian Charter of Rights and Freedoms* in 1982, under which the freedom is guaranteed.[4] In 1985, the Supreme Court of Canada described the 'essence' of this freedom as 'the right to entertain such religious beliefs as a person chooses, the right to declare religious beliefs openly

[1] Ravi Amarnath & Brian Bird is an Associate Lawyer at Blakes, Cassels & Graydon LLP in Toronto, Canada.

[2] Brian Bird is a doctoral student in law at McGill University in Montréal, Canada.

[3] *Mouvement laïque québécois v Saguenay (City)*, 2015 SCC 16, [2015] 2 SCR 3.

[4] *Canadian Charter of Rights and Freedoms*, Part I of the *Constitution Act 1982*, being Schedule B to the *Canada Act 1982* (UK), 1982, c 11, s 2(a) (*Canadian Charter*). It is important to note that s 2(a) of the *Canadian Charter* guarantees 'freedom of *conscience and religion*' (emphasis added).

and without fear of hindrance or reprisal, and the right to manifest religious belief by worship and practice or by teaching and dissemination.'[5] In the same decision, the Court also confirmed that freedom of religion 'equally' protects 'expressions and manifestations of religious non-belief and refusals to participate in religious practice.'[6]

In 2004, the Court defined 'religion' for the purposes of 'freedom of religion' and clarified the test for determining whether the state has infringed this freedom. On the definition of 'religion', the majority of the Court said this:

> Defined broadly, religion typically involves a particular and comprehensive system of faith and worship. Religion also tends to involve the belief in a divine, superhuman or controlling power. In essence, religion is about freely and deeply held personal convictions or beliefs connected to an individual's spiritual faith and integrally linked to one's self-definition and spiritual fulfilment, the practices of which allow individuals to foster a connection with the divine or with the subject or object of that spiritual faith.[7]

As for the legal test for whether freedom of religion is breached, the complainant's (1) belief must be sincere and (2) ability to act in accordance with the belief must have been interfered with in a manner that is more than trivial or insubstantial.[8]

In *Saguenay*, the Court noted that neither the *Canadian Charter* nor the Quebec *Charter of human rights and freedoms*,[9] both of which guarantee freedom of religion, 'expressly imposes a duty of religious neutrality on the state'; the Court concluded, however, that this duty 'results from an evolving interpretation of freedom of conscience and religion.'[10] The Court found evidence of this evolution in a

[5] *R v Big M Drug Mart Ltd* [1985] 1 SCR 295, 18 DLR (4th) 321, 336.

[6] ibid 347.

[7] *Syndicat Northcrest v Amselem*, 2004 SCC 47, [2004] 2 SCR 551 [39].

[8] ibid [56], [59].

[9] *Charter of Human Rights and Freedoms*, CQLR c C-12, s 3 (*Quebec Charter*). The *Quebec Charter*, like the *Canadian Charter*, protects 'freedom of religion'. Section 3 of the *Quebec Charter* also protects freedom of 'conscience', 'opinion', 'expression', 'peaceful assembly', and 'association'.

[10] *Saguenay* (n 3) [71]. The *Canadian Charter* forms part of the Canadian Constitution, which is the supreme law of Canada. The *Quebec Charter* is a provincial statute enacted by the National Assembly of Quebec. It must, like all non-constitutional laws enacted in Canada, conform to the Constitution. Notably, the *Quebec Charter* goes far beyond the content of other provincial human rights legislation in Canada in a number of ways. The most relevant distinction, for the purposes of this comment, is that the *Quebec Charter* guarantees fundamental freedoms such as freedom of religion (in section 3) whereas almost all other provincial human rights legislation in Canada does not.

dissenting opinion within a 2004 decision.[11] The Court concluded in *Saguenay* that the duty of neutrality means that the state can 'neither favour nor hinder any particular belief, and the same holds true for non-belief'.[12]

With these principles in mind, we now turn to the facts and judicial history of *Saguenay*.

3. *Saguenay*—Facts and Judicial History

In *Saguenay*, the Supreme Court of Canada held that freedom of conscience and religion, protected by both the *Canadian Charter* and the *Quebec Charter*, prohibited a municipal council of the City of Saguenay (the 'City') from reciting a Christian prayer at the beginning of its council meetings. Alain Simoneau was a resident of the City, which is in the Canadian province of Quebec. Mr Simoneau, who considered himself an atheist, regularly attended municipal council meetings.

From 2002 to November 2008, the mayor of the City, Jean Tremblay, would commence the City's council meetings by reciting the following prayer:

> O God, eternal and almighty, from Whom all power and wisdom flow, we are assembled here in Your presence to ensure the good of our city and its prosperity.
> We beseech You to grant us the enlightenment and energy necessary for our deliberations to promote the honour and glory of Your holy name and the spiritual and material [well-being] of our city.
> Amen.[13]

Prior to, and after reciting the prayer, the mayor would make the sign of the cross and state: '[i]n the name of the Father, the Son and the Holy Spirit'.[14] Other councillors and municipal officials would cross themselves at the beginning and end of the prayer as well.[15]

Mr Simoneau objected to the prayer, as well as the display of religious symbols—such as a crucifix and a sacred statue—in certain meeting halls. With the support of Mouvement laïque québécois ('MLQ'), a non-profit organisation of which he is a member and which wishes the province to be void of public religious

[11] *Saguenay* (n 3) [71]. See *Congrégation des témoins de Jéhovah de St-Jérôme-Lafontaine v. Lafontaine (Village)*, 2004 SCC 48, [2004] 2 SCR 650 [66]–[67] (LeBel J). Curiously, the Court did not cite *Chaput v Romain* [1955] SCR 834, 1 DLR (2d) 241. *Chaput* predates the *Canadian Charter* but stands for the principle that in Canada there is no state religion, no person must adhere to any religious belief, and all religions are on an equal footing.

[12] *Saguenay* (n 3) [72].

[13] ibid [7].

[14] ibid [6].

[15] ibid.

expression or identity, Mr Simoneau first filed a complaint to the Commission des droits de la personne et des droits de la jeunesse (the 'Commission'), a Quebec government agency which investigates human rights complaints. The Commission informed Mr Simoneau there was adequate evidence for him to bring forth a human rights claim with respect to the prayer. With the continued support of the MLQ, he subsequently filed a complaint to the Quebec Human Rights Tribunal (the 'Tribunal'), including his objections to both the prayer and religious symbols in his complaint. In response, on November 3, 2008, the City adopted By-law VS-R-2008-40 (the 'By-law'), which regulated the prayer's recitation. Specifically, the By-law provided for a two minute delay between the end of the prayer and the official opening of council meetings in order to allow individuals to recuse themselves from the council chamber during the recitation of the prayer.[16] The By-law also provided for a revised prayer, which read as follows:

> *Almighty God, we thank You for the great blessings that You have given to Saguenay and its citizens, including freedom, opportunities for development and peace. Guide us in our deliberations as City Council members and help us to be aware of our duties and responsibilities. Grant us the wisdom, knowledge and understanding to allow us to preserve the benefits enjoyed by our City for all to enjoy and so that we may make wise decisions.*
> *Amen.*[17]

Although the prayer was revised, the mayor and City councillors continued to act in the same way as beforehand (e.g., making the sign of the cross). Consequently, Mr Simoneau and MLQ amended their motion to ask the Tribunal to declare the By-law to be inoperative and of no force or effect in relation to Mr Simoneau.[18]

A. The Tribunal

The Tribunal held that the By-law breached, *inter alia,* section 3 of the *Quebec Charter*, and therefore declared it to be inoperative and of no force or effect.[19] Notably, the Tribunal held that the council had breached its legislative duty to remain neutral between all religions. It stated:

> As the Tribunal explained earlier, when the state and public authorities are involved, they have a duty of neutrality so that the religious equality of everyone is preserved. Considering the conclusions to which the analysis of the religious nature of the

[16] ibid [12].
[17] ibid (emphasis in original).
[18] ibid [13].
[19] *Simoneau c Tremblay*, 2011 QCTDP 1, [2011] RJQ 507.

> prayer and symbols lead, the Tribunal believes that the use of public power to display, in fact convey, a particular faith imposes religious values, beliefs and practices on people who do not share them. In doing so, Ville de Saguenay and the mayor favour one religion to the detriment of another, whereas, pursuant to its duty of neutrality, the state must refrain from intervening to exercise a preference.[20]

Based on the City's failure to abide by its duty of neutrality, the Tribunal concluded that Mr Simoneau could not exercise his rights as an atheist, which section 3 of the *Quebec Charter* also protects.[21]

B. Quebec Court of Appeal

The Quebec Court of Appeal reversed the decision of the Tribunal, holding the Tribunal did not have jurisdiction to adjudicate on the religious symbols and that the prayer did not violate section 3 of the *Quebec Charter*.[22] The Court of Appeal's conclusion on jurisdiction stemmed from the explicit decision of the Commission to restrict its investigation to the prayer. Despite the Commission's decision, the Tribunal concluded that it had jurisdiction to adjudicate both the religious symbols and the prayer.[23]

With respect to the state's duty of neutrality, Gagnon JA, for the majority (and whose reasons were substantially agreed with by Hilton JA), stated:

> I am inclined, for the purposes of this appeal, to adopt the concept of 'benevolent neutrality' used by the author José Woehrling to attempt to better define the parameters of the State's duty of religious neutrality. According to this author, benevolent neutrality is expressed by the respect of all religions, placed on equal footing, without either encouraging or discouraging any form of religious or moral conviction relating directly or indirectly to atheism or agnosticism.[24]

Applying this understanding of neutrality, Gagnon JA noted the historical context of religious symbols and expressions in political institutions throughout Canada, Quebec, and in the City. He concluded: 'A reasonable, well-informed person, aware of the implicit values that underlie this concept could not, in this case, accept the

[20] ibid [250].
[21] ibid [257], [270].
[22] *Saguenay (Ville de) c Mouvement laïque québécois*, 2013 QCCA 936, [2013] RJQ 897.
[23] *Saguenay* (n 3) [10], [14].
[24] *Saguenay* (n 22) [76].

notion that the City's political activities were, because of this prayer, under any particular religious influence.'[25]

C. Supreme Court of Canada

The Supreme Court agreed with the Court of Appeal's jurisdictional conclusion with respect to the religious symbols but restored the Tribunal's decision with respect to the prayer. Gascon J, for the Court, held that the By-law interfered in a discriminatory manner with Mr Simoneau's freedom of conscience and religion protected under the *Quebec Charter*, and that the City's recitation of the prayer contravened the state's duty of neutrality by endorsing one religious tradition. The substance of the duty of neutrality, according to Gascon J, entails that 'the state may not profess, adopt or favour one belief to the exclusion of all others.'[26] Regarding the By-law, he held:

> In a case such as this, the practice of reciting the prayer and the By-law that regulates it result in the exclusion of Mr. Simoneau on the basis of a listed ground, namely religion. That exclusion impairs his right to full and equal exercise of his freedom of conscience and religion. The discrimination of which he complains relates directly to the determination of whether, on the one hand, the prayer is religious in nature and whether, on the other hand, the City is entitled to have it recited as it did.[27]

The Supreme Court agreed with the Tribunal's conclusions that freedom of conscience and religion under the *Quebec Charter* protected the 'freedom not to believe, to manifest one's non-belief and to refuse to participate in observance' and that the prayers recited at the City's meetings had a religious purpose.[28] Accordingly, the Court held that the City breached its 'duty of neutrality', which 'requires that the state neither favour nor hinder any religion, and that it abstain from taking any position on this subject.'[29]

[25] ibid [107].
[26] *Saguenay* (n 3) [84].
[27] ibid [64].
[28] ibid [70], [114].
[29] ibid [137].

4. Implications of *Saguenay*

Saguenay directs that the Canadian state cannot favour one religion over others—it must be neutral on the matter of religion in carrying out its functions. This does not mean that the state is allowed to explicitly favour unbelief to belief. Yet, as Gascon J admitted in *Saguenay*, the difference between doing so and not doing so is 'subtle'.[30] Among these subtleties, we believe the Court's interpretation of the duty of neutrality gives rise to three interesting issues.

A. What is the Distinction between Absolute Neutrality and True Neutrality?

Saguenay raises the issue of what distinguishes 'absolute' and 'true' neutrality. The respondents—the City and its mayor, Jean Tremblay—argued that barring the council's prayer amounted to the state preferring atheism and agnosticism to theistic religious belief.[31] While Gascon J readily accepted the difficulty in achieving religious neutrality in the public square, he rejected the respondents' argument:

> [A]bstaining does not amount to taking a stand in favour of atheism or agnosticism. The difference, which, although subtle, is important, can be illustrated easily. A practice according to which a municipality's officials, rather than reciting a prayer, solemnly declared that the council's deliberations were based on a denial of God would be just as unacceptable. The state's duty of neutrality would preclude such a position, the effect of which would be to exclude all those who believe in the existence of a deity.
>
> In short, there is a distinction between unbelief and true neutrality. True neutrality presupposes abstention, but it does not amount to a stand favouring one view over another. No such inference can be drawn from the state's silence.[32]

Gascon J is correct that a council meeting featuring an express denial of God would infringe the state's duty of neutrality. Yet even if the City attempted to respect all religious traditions to which its population adheres (for example, by rotating through prayers and spiritual readings from these traditions), this would still infringe the religious freedom of the atheist or agnostic. Indeed, Gascon J

[30] ibid [133].
[31] ibid [130].
[32] ibid [133]–[134].

confirmed that even a non-denominational prayer would violate the duty of neutrality because it still excludes non-believers.[33]

It is unclear in *Saguenay* whether the Court views atheism as a religion or religious belief. If it does, then cases like *Saguenay* could be viewed as cases of competing religions (i.e. atheism vs Christianity). If this is the case, it is arguable that prohibiting the state from showing traditional religious symbols or reciting prayers is not a matter of enforcing a 'duty of neutrality' but rather amounts to the state favouring one religion over another.

If the only way for the state to fulfil the duty of neutrality is by not professing any religious view at any moment (and assuming atheism and agnosticism are not religions), then it follows that the duty favours atheism and agnosticism in the public square. The example given by Gascon J seems to be a distinction in degree rather than in kind. This is to say that allowance for the state to deny God's existence is simply a greater (or more obvious) state preference for atheism and agnosticism than the duty of neutrality as defined in *Saguenay*. It is difficult to identify the distinction between the duty of neutrality as defined in *Saguenay* and a duty of irreligion on the part of the state. Regardless of its purpose, the unavoidable effect of the duty of neutrality is to favour atheism or agnosticism over theistic religion in the public square.

B. Are there Exceptions to the Duty of Neutrality?

After *Saguenay*, must there be no prayers recited at any municipal council meeting in Canada? The Court did not answer this question explicitly, but the decision appears to favour prayer-free meetings. But what if all of the attendees at a council meeting consent to the recitation of a prayer? And what if all of the attendees adhere to the same religion and a prayer from that religious tradition is recited? *Saguenay* does not opine on these scenarios. If a council meeting begins with a prayer in one of these scenarios, this would certainly violate the state's duty of neutrality because the state, by allowing the prayer, has preferred one religion to another.

The question here is whether there are exceptions to the duty of neutrality. In our view the duty of neutrality is intended to protect the rights of citizens vis-à-vis the state. The trigger for the *Saguenay* litigation was Mr Simoneau's individual right under the *Quebec Charter* to not adhere to any religion. Had Mr Simoneau been undisturbed by the prayer's recitation and not challenged this state practice, there would appear to be no legal bar to the recitation of the prayer.

Saguenay also does not squarely engage with the religious freedom of the individual who is a state official (e.g., the mayor of Saguenay, Jean Tremblay). The Court focused on the religious freedom of Mr Simoneau. Must a person who is a state official leave their religious identity at the door of their workplace as they

[33] ibid [137].

do their coat? Going forward, can Mr Tremblay silently pray while seated in the council chamber, make the sign of the cross, and start the meeting? This may also cause some discomfort for Mr Simoneau, but it would seem heavy-handed to deny state officials the right to express their religious identity in this way—especially where the intent of the expression is to seek assistance in the performance of their duties. It would be peculiar, indeed, to refuse Mr Tremblay the right to silently pray that he be effective and competent in his service of the citizens of Saguenay, Mr Simoneau included.

Before turning to the relationship between the duty of neutrality and secularism, we note that *Saguenay* may identify an exception to the duty of neutrality. At one point in the decision, Gascon J referred vaguely to the 'many traditional and heritage practices' in Canada 'that are religious in nature.'[34] Without providing specific examples, Gascon J held that 'it is clear that not all of these cultural expressions are in breach of the state's duty of neutrality'.[35]

In our view it is difficult to identify 'traditional and heritage practices' that are religious that do not breach the state's duty of neutrality. Indeed the category appears quite narrow, as Gascon J stated that if the traditional or heritage practice reveals 'an intention to profess, adopt or favour one belief to the exclusion of all others, and if the practice at issue interferes with the freedom of conscience and religion of one or more individuals, it must be concluded that the state has breached its duty of religious neutrality.'[36]

We cannot presently think of a religiously inspired traditional or heritage state practice in Canada that would respect the duty of neutrality in light of this test. Indeed, while *obiter dicta*, Gagnon JA of the Quebec Court of Appeal viewed the council's display of religious symbols, which could be construed as professing, adopting or favouring one belief to the exclusion of all others, as referencing Quebec's 'cultural and historical heritage.'[37] Yet the result in *Saguenay* appears to challenge this conclusion. On this front the Court seems to be trying to put some of the toothpaste back in the tube after letting it out.

C. What about Secularism?

Saguenay does not explicitly discuss the relationship between the duty of neutrality and secularism. Secularism and related words like 'secularity' scarcely appear in the decision. Yet Canada's identity as a secular state rests just beneath the surface in *Saguenay*. The duty of neutrality discussed in *Saguenay* is concerned with how

[34] ibid [87].
[35] ibid.
[36] ibid [88].
[37] *Saguenay* (n 22) [156].

secularism in Canada is to be achieved, but there is no direct discussion of the *kind* of secularism that Canada is pursuing.

In *Secularism and Freedom of Conscience*, Jocelyn Maclure and Charles Taylor argue that secularism rests on two principles (equality of respect and freedom of conscience) and on two 'operative modes' that make it possible to achieve these principles (separation of church and state and state neutrality toward religions).[38] The two principles are the *ends* of secularism; the operative modes are the *means*. How these principles and operative modes are interpreted and applied leads to differing versions of secularism that are more or less restrictive on *individual*—in addition to *state*—expression of religion. Maclure and Taylor describe two dominant versions of secularism: the (1) 'republican' model which 'allows greater restriction on the free exercise of religion, in the name of a certain understanding of the state's neutrality and of the separation of political and religious powers' and the (2) 'liberal-pluralist' model 'centered on the protection of freedom of conscience and of religion, as well as a more flexible concept of separation and neutrality.'[39] The republican view demands neutrality from individuals in public (to varying degrees) while the liberal-pluralist view requires neutrality of institutions but not individuals.[40]

In our view, *Saguenay* supports the liberal-pluralist version of secularism. For example, Gascon J stated:

> By expressing no preference, the state ensures that it preserves a neutral public space that is free of discrimination and in which true freedom to believe or not to believe is enjoyed by everyone equally, given that everyone is valued equally. *I note that a neutral public space does not mean the homogenisation of private players in that space. Neutrality is required of institutions and the state, not individuals* (…) . On the contrary, a neutral public space free from coercion, pressure and judgment on the part of public authorities in matters of spirituality is intended to protect every person's freedom and dignity. (…).
>
> I would add that, in addition to its role in promoting diversity and multiculturalism, the state's duty of religious neutrality is based on a democratic imperative. The rights and freedoms set out in the *Quebec Charter* and the *Canadian Charter* reflect the pursuit of an ideal: a free and democratic society. This pursuit requires the state to encourage everyone to participate freely in public life regardless of their beliefs (…) . The state may not

[38] Jocelyn Maclure and Charles Taylor, *Secularism and Freedom of Conscience* (Jane Marie Todd tr, Harvard University Press 2011) ch 2.

[39] Maclure and Taylor (n 38) 27.

[40] ibid 39.

> act in such a way as to create a preferential public space that favours certain religious groups and is hostile to others. *It follows that the state may not, by expressing its own religious preference, promote the participation of believers to the exclusion of non-believers or vice versa.*[41]

Gascon J also held in *Saguenay* that 'what is in issue here is not complete secularity, but true neutrality on the state's part and the discrimination that results from a violation of that neutrality.'[42] It appears from *Saguenay* that the Canadian Constitution does not envision a Canada in which religious expression must be absent from the public square with respect to *individual* expression—even where the individual is working as a state official.

This sentiment was buttressed by another Supreme Court of Canada case concerning freedom of religion decided one month prior to *Saguenay*, where the majority of the Court held:

> Part of secularism, however, is respect for religious differences. A secular state does not—and cannot—interfere with the beliefs or practices of a religious group unless they conflict with or harm overriding public interests. Nor can a secular state support or prefer the practices of one group over those of another: (…). The pursuit of secular values means respecting the right to hold and manifest different religious beliefs. A secular state respects religious differences, it does not seek to extinguish them. Through this form of neutrality, the state affirms and recognises the religious freedom of individuals and their communities.[43]

These sentiments emphasise that, in Canada, religion is not to be totally void in the public sphere. There is a material difference between allowing a Christian government employee to say grace before eating lunch at the workplace and having the Christian prayer recited at the beginning of the lunch hour over the public announcement system for all employees to hear. It is important that the holding in *Saguenay*, which concerns the state's duty of neutrality regarding religion and irreligion, not be construed as pertaining to individual religious or irreligious expression in public.

[41] *Saguenay* (n 3) [74]–[75] (emphasis added).
[42] ibid [78].
[43] *Loyola High School v Quebec (Attorney General)*, 2015 SCC 12, [2015] 1 SCR 613, [43]–[44].

5. Conclusion

The primary contribution of *Saguenay* is the principle that the Canadian state bears a duty of neutrality with respect to religion (and irreligion)—a duty that 'results from an evolving interpretation of freedom of conscience and religion' in Canadian constitutional law.[44] As we have explored, the imposition of this duty raises difficult issues and leaves important questions unanswered for the time being. These issues and questions include the distinction between absolute and true neutrality, whether true neutrality favours unbelief to belief in the public square, and determining which Canadian 'traditional and heritage practices' connected to religion can continue (if any). With respect to these matters, only time—and further litigation—will bring answers.

At the same time, *Saguenay* provides guidance with respect to the version of secularism that the Canadian Constitution envisions, namely a version that does not render the public square a 'religion-free' zone with respect to individual expression. It is ironic that *Saguenay* originated in Quebec because, in 2013, the government of that province proposed a law—the so-called 'Secular Charter of Values'—which would have prohibited public servants from wearing certain types of religious apparel at work. It was expected that the Secular Charter would apply to articles such as the Sikh turban, Jewish kippah, and Muslim hijab, among others.[45] The proposed law—which died on the Order paper by virtue of a provincial election—caused great controversy throughout Canada and sparked heated debate on the version of secularism that Canada should pursue. *Saguenay*, despite barely mentioning secularism, hints rather strongly that this proposed law subscribes to a version of secularism that is incompatible with the 'free and democratic society' that the *Canadian Charter* envisions for Canada.[46] In this respect, *Saguenay* may feature 'more than meets the eye' in terms of the impact it will have on Canadian constitutional law—and Canada itself.

[44] *Saguenay* (n 3) [71].

[45] Bill 60, *Charter affirming the values of State secularism and religious neutrality and of equality between women and men, and providing a framework for accommodation requests*, 1st Sess, 40th Leg, Quebec, 2013. Section 5 of Bill 60 prohibited public servants in Quebec from wearing 'objects such as headgear, clothing, jewelry, or other adornments which, by their conspicuous nature, overtly indicate a religious affiliation.'

[46] *Canadian Charter* (n 4) s 1.