CAMBRIDGE LAW REVIEW

Volume VI
Issue 1
Spring 2021

Editors-In-Chief
Despoina Georgiou

Proudly Supported By
Cambridge University Law Society

Editors-in-Chief's Introduction to the Spring Issue of Volume V of the Cambridge Law Review

It is with great pleasure that I present the Spring Issue of Volume VI of the Cambridge Law Review. Thanks to the remarkable quality of the submissions and editing, the journal has managed, in just a few years' time, to become recognised as a high-calibre publication. This year has been our busiest one yet. We strengthened our previously established partnerships with the Oxford Undergraduate Law Journal and the London School of Economics Law Review and we also built new relationships. As of 2021, the Cambridge Law Review is proud to be partners with the Bristol Law Review, the Exeter Law Review, the Durham Law Review, and the Harvard Undergraduate Law Review. We also participated in educational seminars where we discussed various aspects of academic publishing with Editors-in-Chief of law reviews from around the world (such as the University of Bologna Law Review, the Auckland University Law Review, and the Oxford Undergraduate Law Journal).

The increased exposure combined with the interest generated by the high-quality of the articles published in the previous Volumes raised the journal's profile, leading to a record number of submissions for the Spring Issue of Volume VI. For this reason, we decided to double the number of articles published in this Issue. Volume VI, Issue I comprises scholarship from a variety of disciplines. The articles published deal with contemporary matters in the areas of Constitutional Law, Indigenous Law, International Arbitration, Financial Services Regulation, Company Law, International Human Rights Law, Discrimination Law, Tort Law, and others. More particularly:

In his article "Constitutional Courts' Activism and the Relation Between Law and Politics: A Legal Theoretical Contribution", Professor Mauro Zamboni examines the role and place of Constitutional Courts in Western or Western-like democracies. As he argues, even though Constitutional Courts play a bridging role between the political and legal worlds, they are – from an institutional and functional

perspective – primarily legal actors. Therefore, their position as institutional actors should be based upon the direct effects of their decisions ('outputs') in the legal arena, rather than on the indirect consequences ('outcomes') in the political arena. The article concludes that the Constitutional Courts' primary responsibility ought to be towards the legal community and the paradigms governing its discourse.

Professor Frankie Young writes in the area of Indigenous Law and Private International Law (Conflict of Laws). His article "Positioning Indigenous Law in the Legally Pluralistic State of Canada" constitutes a commentary on the Beaver v Hill judgement. This is a key legal decision from the State of Canada that deals with the application of Private International Law to resolving a Family Law dispute involving indigenous litigants. Providing a well-reasoned analysis of the Court of Appeal's judgement, Young sheds light on contentious issues regarding the application of Indigenous Law.

Domenico Piers De Martino and Dr Katharina Plavec write on the topical issue of 'digital arbitration'. Their article "Has COVID-19 Unlocked Digital Justice? Answers from the World of International Arbitration" presents the legal framework regarding online hearings and examines how arbitral institutions around the world have adapted to the constraints imposed by the COVID-19 pandemic. After analysing the relevant regime, the authors conclude that a single, uniform and exhaustive answer on the legality of virtual hearings is not possible. This is because the answer is conditional upon the position adopted in legislation across multiple jurisdictions and requires an ad hoc approach. Nevertheless, they find that, in general, remote hearings are permissible under the New York Convention, and are not prohibited by the national arbitration laws of the analysed jurisdictions. Therefore, they predict that remote hearings will be more widely adopted in the near future.

Aleksander Kalisz also writes in the area of International Arbitration ("Illegal and Inappropriate Evidence in International Investment Law: Balancing Admissibility"). Given that no clear test has been laid down in the applicable procedural rules or treaties regarding the admissibility of illegally or inappropriately obtained evidence, Kalisz uses case law to examine whether a common test for admissibility can be inferred from arbitral decisions. Case law, in this context, is relevant because, although there is no doctrine of precedent in Investment Law, tribunals are prompted to follow a harmonious interpretation of International Law and previous cases are deemed highly authoritative. Besides case law, the article examines the procedural principles enshrined in Bilateral Investment Treaties (BITs), arbitration rules, and rules on the taking of evidence. Particular emphasis is paid to the International Centre for Settlement of Investment Disputes (ICSID) Convention and Arbitration Rules and the United Nations Commission

on International Trade Law (UNCITRAL) Arbitration Rules. Other non-binding instruments (2020 International Bar Association Rules on the Taking of Evidence, 2018 Rules on the Efficient Conduct of Proceedings in International Arbitration) are also examined to provide a full picture of the legal regime in place.

In her article "Reimagining a Centralised Cryptocurrency Regulation in the US: Looking Through the Lens of Cryptoderivatives", Sangita Gazi presents a comparative analysis of the US regulatory responses to crypto-derivatives with specific references to the UK's and the EU's approaches and rationale towards crypto-derivatives regulations in their respective regions. Through a well-reasoned analysis, Gazi argues that it is paramount that the US enacts comprehensive cryptocurrency regulation that recognizes the novelty of cryptocurrencies' market risks and introduces a robust regulatory infrastructure to limit market manipulation in the cryptocurrency spot market vis-à-vis the crypto-derivatives market. Gazi envisions a cryptocurrency regulation that includes: (i) a centralised cryptocurrency trading platform; (ii) a mandatory registration requirement for all cryptocurrency exchanges and; (iii) a federal cryptocurrency agency. She suggests that, with a degree of centralisation, a federal cryptocurrency agency is likely to establish the desired visibility into the cryptocurrency spot and an effective oversight mechanism that would eventually help curb market manipulation and restore investor confidence.

Mikołaj Kudliński writes in the area of Company Law. His article "Are Involuntary Creditors Adequately Protected from the Adverse Impact of the Doctrine of Limited Liability? An Analysis of the Origins of the Doctrine and its Modern Application Through the Prism of Involuntary Creditors' Protection", discusses the origins of the limited liability doctrine in the UK law. As it finds, the interests of involuntary creditors were not given adequate consideration at the time of its inception, with the doctrine not being conceptualised to apply to this group of creditors at all. The article analyses the current protection mechanisms available to creditors and discusses alternative approaches to limited liability. As it argues, a control-based presumption of parent liability would strike a fair balance between the interests of the various actors involved in the company's activity, providing involuntary creditors with a greater degree of protection.

Mohamed El Eryan writes on the contentious topic of Iraqi Kurdish self-determination. His paper "Iraqi Kurdish Self-Determination: A Pathway to Secession? Settling the Questions of Application & Scope" examines the extent to which Iraqi Kurds are a people with a right to self-determination and assesses whether that right can express itself through remedial secession. El Eryan finds that there is insufficient support for the existence of a positive right to remedial secession and argues that, even if such a right existed or was to develop in the

future, the situation in Iraqi Kurdistan would not meet the high threshold required for remedial secession to be triggered. For this reason, El Eryan suggests that a political solution based on a broader autonomy arrangement and increased forms of cooperation is needed to resolve the continuing disputes between the Iraqi Federal Government and Iraqi Kurdistan. As he says, until Iraqi Kurds can rely on regional and external political frameworks that provide the required support for statehood, a Kurdish state will not be viable.

In her article "Marking the Internal and External Limits of Discrimination Law in Lee v Ashers Baking Company", Emily Mei Li Ho comments on the UK Supreme Court's Lee v Ashers Baking Company decision. The case involved bakers who refused to fulfil a customer's order of a cake iced with the message 'Support Gay Marriage'. The UK Supreme Court decided in favour of the bakers, and in so doing, analysed and marked the limits of discrimination law – specifically, the prohibition of direct discrimination. In her article, Ho marks these limits, examining their desirability against the background of domestic and international jurisprudence and political theory concerning freedoms of religion and expression. She concludes that the decision was a welcome bridling of discrimination law – an area in which expansions can be tempting owing to the nobility of the aim of equality – but which must be limited for the sake of other liberal values.

Nicholas Goldrosen writes in the area of Criminal Law. In his article "What Happens in the Jury Room Stays in the Jury Room: R v Mirza, the Criminal Justice and Courts Act, and the Problem of Racial Bias", Goldrosen argues that the courts' refusal to consider juror testimony about deliberations and the laws restricting jurors from speaking about deliberations prevent defendants from seeking adequate redress for juror racial bias. As exemplified in the R v Mirza decision, English courts have historically upheld jury secrecy by holding that the interests of finality and candour outweigh the injury done to a defendant by juror racial bias. While the Criminal Justice and Courts Act 2015 has introduced some changes to jury secrecy law (mainly by allowing jurors to report some forms of misconduct that occur during deliberations), these are not adequate in protecting defendants' rights. As Goldrosen shows, the Act's reporting provisions are overly complex, largely non-adversarial, and too focused on enabling the prosecution of jurors who commit misconduct. For this reason, the author argues that a reform of this Act to more explicitly focus on protecting defendants from juror misconduct – and in particular, juror racial bias – is necessary to better secure defendants' fair trial rights.

The last article of this issue is written by Soh Kian Peng ("Spandeck: A Relational View of the Duty of Care"). Relying on the example of the Spandeck framework in Singaporean jurisprudence, Peng presents the argument that such

frameworks – being consistent with a relational conception of tort law– can provide a useful means of determining whether a duty of care exists. In so doing, the article addresses some criticisms of the relational view and re-emphasises the important role the duty of care plays in the tort of negligence.

Overall, the ten articles included in this Issue constitute exceptional pieces of academic work that enrich the literature in their respective fields. They provide valuable insights into the selected areas of research, constituting enjoyable reads that would be of interest to British and international, academic and professional audiences alike.

I owe heartfelt thanks to our team of Associate, Senior, and International Editors for their dedication and work during these challenging times. Despite the difficulties caused by the COVID-19 pandemic and the subsequent lockdowns, the Editorial Board worked tirelessly to ensure the highest standards of quality for this Issue. I would also like to express my gratitude to the Honorary Board for their invaluable guidance and to the Cambridge University Law Society for their continued support, without which this Issue would not have been possible. I look forward to presenting the Autumn Issue which will be published later in the year.

Despoina Georgiou
Editor-in-Chief

Cambridge Law Review

Table of Contents

Cambridge Law Review (2021) Vol VI, Issue i, 1–29

Constitutional Courts' Activism and the Relation Between Law and Politics: A Legal Theoretical Contribution

Mauro Zamboni*

Abstract

If one looks at the Constitutional Courts and their place in the architectural landscape of a constitution in a Western or Western-like democracy, they appear to lean into politics through their activism in the political game. This leaning appears to go against the (at least theoretical) nature of a court, which has the task of settling legal cases, not dealing with political decisions. In particular, many Constitutional Courts in the Western or Western-like democracies have grown more distant from the legal arena (as the ultimate authority on what constitutes valid law) and closer to the political one (by limiting and expanding the operating space of the political actors and their legislative tools). The basic thesis of this article is that, from a descriptive point of view, Constitutional Courts, though playing a bridging role between the political and legal worlds, are – from an institutional and functional perspective – primarily legal actors. Without a doubt, these Courts play a role in the political game; however, their position as institutional actors should be based upon the direct effects of their decisions ('outputs') in the legal arena, rather than on the indirect consequences ('outcomes') in the political arena. As a consequence, the article comes to the conclusion that the Constitutional Courts'

* Professor in Legal Theory at the Faculty of Law, Stockholm University, Sweden. I am grateful to Jane Reichel, Maria Grahn Farley, Liane Colonna, and the anonymous reviewers for their very helpful comments on earlier drafts. Any errors that remain are my own. Mauro.Zamboni@juridicum.su.se.

primary responsibility ought to be towards the legal community and the paradigms governing its discourse.

Keywords: constitutional courts, politics, judicial activism, functions, institutional position

I. Introduction

If one looks at the Constitutional Courts and their place in the architectural landscape of a constitution in a Western or Western-like democracy, they appear just like the Leaning Tower of Pisa: they are essential in giving the landscape a certain character, they are beautiful to behold, but they also instill a certain discomfort. Just as the Tower of Pisa has an obvious feature that is quite unnatural for a tower, that is, deviating from the 'natural' position of standing straight in relation to the ground, so the Constitutional Courts, when observed in their constitutional landscapes, can immediately be seen to lean into politics through their activism in the political game. This leaning appears to go against the (at least theoretical) nature of a court, which has the task of settling legal cases, not dealing with political decisions.

Such leaning towards the political game on the part of the judges (in particular at the highest levels) is not a novelty of our times, going back at least to the late 18th century (as seen in the role of judges in the French Revolutionary state) or the early 19th century (as seen in the role of the Supreme Court in the institutional building of the United States).[1] However, the role that the judicial bodies play in the political arena has become one of hottest topics of debate in the legal world in the last decades, often under the label of "judicial activism" or "judicialisation of politics".[2] As regards what are usually the highest of the judicial bodies, namely the Constitutional Courts (or under any other court operating as a guardian of the respect for the constitutional documents on the part of the other

[1] E.g., the French Loi des 16-24 août 1790 sur l'organisation judiciaire, Title II *Article* 12 <www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000704777> *accessed 14 March 2021*; or the famous *Marbury v Madison* 5 US (1 Cranch) 137 (1803).

[2] Ran Hirschl, 'The Judicialization of Politics' in Robert E Goodin (ed), *The Oxford Handbook of Political Science* (OUP 2011) 253. E.g., Anke Eilers, 'The Binding Effect of Federal Constitutional Court Decisions upon Political Institutions' (European Commission for Democracy Through Law (Venice Commission, 22 May 2003) <www.venice.coe.int/docs/2003/CDL-JU(2003)018-e.pdf> accessed 14 March 2021; or Sandra Day O'Connor, 'On the Importance of Having A Fair and Independent Judiciary' (2006) 38 Third Branch 9, 10 ("Directing anger toward judges has had a long tradition in our nation … While scorn for some judges is not altogether new, I do think that the breadth of the unhappiness being currently expressed, not only by public officials but in public opinion polls in the nation, shows that there is a level of unhappiness today that perhaps is greater than in the past and is certainly cause for great concern.").

legal actors and the law-makers), one can easily see a trend in the recent decades. For reasons which will not be explored in this article, many Constitutional Courts in the Western or Western-like democracies have become increasingly 'active' by furthering their activities in a law-making direction and offering authoritative interpretations of the valid law, different from that of other actors at the top of the constitutional structure which are institutionally devoted to law-making (e.g., the national assemblies or the executive branch).[3] In other words, many Constitutional Courts have shifted their 'recognised' role of being the authoritative interpreter of the law, in light of the constitutional documents, towards the middle of the institutional map of a democracy based on the rule of law. Thus, they have grown more distant from the legal arena (as the ultimate authority on what constitutes valid law) and closer to the political one (by limiting and expanding the operating space of the political actors and their legislative tools).[4]

As can be understood from this concise introduction, judicial activism (in particular on the part of the highest courts) focuses primarily on the relations between law and politics. Therefore, the main purpose of this article is to evaluate whether Constitutional Courts should be considered, according to legal theoretical criteria, to be legal actors simply enforcing what is written by political actors in statutes and constitutions, or institutional actors with a predominantly political nature, determining what the law should say. To achieve this, the investigation will not only focus on the function that Constitutional Courts play, between law and politics, but also on their institutional positioning itself.

When reading the legal literature on Constitutional Courts and politics, it is evident that important concepts are applied inconsistently and imprecisely. The result is a befuddlement among scholars whereby, in certain instances, the same concept can have several different interpretations. An example of the terminological inexactitude used in the discourse concerns the word 'political'. When some academics refer to 'political' in the Constitutional Courts, they refer to the fact that these judicial bodies perform the work of politicians. Yet, other scholars use the word to identify the fact that Constitutional Courts are themselves

[3] E.g., Richard A Posner, 'The Supreme Court 2004 Term–Foreword: A Political Court' (2005) 119 Harvard Law Review 32, 35–39; Michael Hein, 'The Least Dangerous Branch? Constitutional Review of Constitutional Amendments in Europe' in Martin Belov (ed), *Courts, Politics and Constitutional Law: Judicialization of Politics and Politicization of the Judiciary* (Routledge 2020) 187–206.

[4] Joseph Raz, 'On the Nature of Law' in Joseph Raz, *Between Authority and Interpretation:* On the Theory of Law and Practical Reason (OUP 2009) 99–100. E.g., Julien Mouchette, 'The French Constitutional Council as a Law-Maker: From Dialogue with the Legislator to the Rewriting of the Law' in Monika Florczak-Wątor (ed), *Judicial Law-Making in European Constitutional Courts* (Routledge 2020) 9–27; Harvie J Wilkinson, 'Of Guns, Abortions, and the Unraveling Rule of Law' (2009) 95 Virginia Law Review 253, 275–288.

political actors, just like any other political party, with their own political agenda. In particular, this article aims to better understand the role of the politics in the work of the Constitutional Courts. It suggests that legal practitioners and scholars must reflect upon and more clearly define, with the help of political and social-science scholarship, where the institutional positioning of these Courts are in the legal architecture of a Western or Western-like democracy and the functions such judicial bodies should play in the political arena.

The basic thesis is that, from a descriptive point of view, Constitutional Courts, though playing a bridging role between the political and legal worlds, are – from an institutional and functional perspective – primarily legal actors. Without a doubt, these Courts play a role in the political game; however, their position as institutional actors should be based upon the direct effects of their decisions ('outputs') on the legal arena, rather than on the indirect consequences ('outcomes') on the political arena. As a consequence, and moving from these descriptive findings into the realm of normative judgements, the article comes to the conclusion that the Constitutional Courts' primary responsibility ought to be towards the legal community and the paradigms governing its discourse.

This article makes no pretensions on providing any final words either in the discussion on judicial activism or as regards the relations between the legal and political worlds in general. First, the focus here is solely on Western (or Western-like) legal systems; second, this work has a more limited objective of contributing to clarifying the terms of the discussion, in particular by finding a solid basis (at least from a legal perspective) upon which to begin the discussion on whether judicial activism is good or bad. In other words, this investigation primarily has a *legal theoretical* task: To clarify what is meant when the legal discussion deals with the 'political' in the work of Constitutional Courts.[5] In particular, the legal theoretical approach is considered an appropriate lens because, if used in HLA Hart's terms, it enables one to consider how the concept 'political' is conceived and employed inside the legal order by legal actors when talking about the Constitutional Courts. This approach also allows one to clarify the specific meaning of the idea of politics in the Constitutional Courts within a legal context, distinguishing it then from the use it can have in ordinary everyday or political language. By offering such "elucidation of the use of words in characteristic legal contexts", the legal theoretical approach can then provide some normative criteria to the legal actors on how to evaluate the work done by such high courts when the Constitutional Courts come in contact with the political arena or, in general, with highly political

[5] HLA Hart, 'Analytical Jurisprudence in Mid-Twentieth Century: A Reply to Professor Bodenheimer' (1957) 105 University of Pennsylvania Law Review 953, 961–962. See also HLA Hart, 'Definition and Theory in Jurisprudence' (1954) 70 LQR 37, 60.

issues during their work.[6] In other words, the legal theoretical approach, by making descriptively clear in what sense 'politics' enters into the work of the Constitutional Courts, helps also to set the stage for a more normative enterprise aimed at indicating to legal actors what ought to be considered (and not) when evaluating the 'political' work done by Constitutional Courts.

In Part I, certain key concepts used in this article are defined. The focus in Part II is on the importance of judicial activism when dealing with Constitutional Courts. Part III moves the attention to a feature of Constitutional Courts in well-established Western democracies, namely their positioning between the legal and political arenas and how this hybrid nature may be a theoretical problem. The final Parts IV and V then identify why the characterisation of Constitutional Courts, as either legal or political actors, is relevant from a legal perspective, both in descriptive and normative terms. These parts also explore how it is possible, at least from a legal theoretical perspective, to resolve the dilemma of 'legal vs. political actors' by defining Constitutional Courts as legal actors performing a political function.

II. A Definitional Framework

Before we delve into the topic at hand, some key concepts used in this text must be clarified. First, *judicial activism*, sometimes referred to by other words (e.g., 'constitutional politics', 'government of judges', or 'judicialisation of politics'), identifies the general phenomenon – typical of (but not limited to) well-established Western democracies – when "courts impose a judicial solution over an issue erstwhile subject to political resolution" by intervening against and striking down a part of properly enacted legislation, or by "legislating" in an area in the absence of legislation.[7] Judicial activism thus identifies a judicial activity directed at stretching the formal structures and the letter of the law to fill gaps (or what are perceived as gaps) left by politicians. The judges take a more 'active' stance towards law-making (in particular in its legislative form) to implement values that the political

[6] Hart, 'Analytical Jurisprudence in Mid-Twentieth Century: A Reply to Professor Bodenheimer' (n *5*). See also HL*A* Hart, *The Concept of Law* (3rd edn, OUP 2012) Chapters II and IV; and Hart, 'Definition and Theory in Jurisprudence' (n 5) 38.

[7] David L Anderson, 'When Restraint Requires Activism: Partisan Gerrymandering and the Status Quo Ante' (1990) 42 Stanford Law Review 1549, 1570. See also Keenan D Kmiec, 'The Origin and Current Meanings of Judicial Activism' (2004) 92 California Law Review 1441, 1463–1476; Alec Stone Sweet, *Governing with Judges: Constitutional Politics in Europe* (OUP 2000) 61–66; Charles S Lopeman, *The Activist Advocate: Policy Making in State Supreme Courts* (Preager 1999) 3. *cf* Frank H Easterbrook, 'Do Liberals and Conservatives Differ in Judicial Activism?' (2002) 73 University of Colorado Law Review 1403, 1403; William P Marshall, 'Conservatives and the Seven Sins of Judicial Activism' (2002) 73 University of Colorado Law Review 1217, 1220–*122*1 (pointing out to the difficulties of identifying a single meaning of "judicial activism").

actors are unable to sense in the community or are unable to transform into legislative measures, or that are part of the political baggage of certain judges.[8] Usually, the courts find support for their new course in the foundational structures of the legal system, for example, the constitution or international treaties. In other words, judicial activism refers to the complex of judicial activities through which the judges consciously and explicitly take upon themselves a power that has traditionally been left to other institutional actors, for example, the political actors sitting in the national assemblies or (to a lesser extent) the public administration. In doing so, the courts are guided by the idea that their primary role is neither to find the true intention of the legislative bodies nor to review the work done by the public agencies. Instead, they intend to act as guardians of the legal system as a whole, by positioning themselves as a third party and solving disputes in the light of fundamental legal principles which have not been contemplated by the legislative bodies and which have been neglected in administrative practice.

Second, when referring to Constitutional Courts, this includes all the highest courts that – under varying names (e.g., High Council or Supreme Court) – have among their primary legal duties the jurisdiction to evaluate the constitutionality of the law, for example, the consistency or conflict of legally relevant documents produced within a certain legal system in relation to the basic legal documents of that community. Such courts are also characterised, at least with respect to conducting constitutional reviews, by being positioned outside the ordinary court system, in a sense, and by their work being completely independent (at least in its modality) from the other branches of the state.[9] It is worth noting that several types of Constitutional Courts can find a place under this definitional umbrella. In particular, this definition allows scrutiny both of Constitutional Courts that have an abstract review competence (e.g., when a Constitutional Court is asked to determine the compatibility of statutory law with the Constitution at the request of non-judicial public bodies, e.g., a law-drafting committee of the national assembly or a regional government), and those that have a more concrete review power

[8] E.g., Aharon Barak, *The Judge in a Democracy* (Princeton University Press 2006) 8–10. See also Niels Petersen, *Proportionality and Judicial Activism:* Fundamental Rights Adjudication in Canada, Germany and South Africa (CUP 2017) 18–33 (pointing out judicial activism as aimed to correct the failures of the political market).

[9] Ralf Rogowski and Thomas Gawron, 'Constitutional Litigation as Dispute Processing' in Ralf Rogowski and Thomas Gawron (eds), *Constitutional Courts in Comparison:* The US Supreme Court and the German Federal Constitutional Court (2nd edn, Berghahn 2016) 1. Due to this very "external" position in relation to the ordinary judiciary, Western legal systems usually assign to Constitutional Courts other fundamental "above the parties" types of jurisdiction. Víctor Ferreres Comella, *Constitutional Courts and Democratic Values: A European Perspective* (Yale University Press 2009) 6.

(e.g., when a Constitutional Court's review jurisdiction is activated by a party to litigation, or by a lower judge, stating that a law violates the constitutional texts).[10]

Third, the definition of 'political actors' adopted here is fairly different from that used by the other discipline that also investigates the role of Constitutional Courts in relation to the political system, namely political science. While for the latter, political actors are more or less identified as all institutional actors that 'make the law', the term in this article is intended from a more legal perspective and refers to a narrower range of institutional entities whose primary goal is to see their values implemented into a community by making use of the legal apparatus and system, for example, political parties or interest groups. Political actors can (and usually do) have a primary goal of a non-legal nature (e.g., an economic or social nature) and therefore, in their operations, mainly take into consideration the environments surrounding the legal arena, for example, the political or socio-economic ones. Moreover, their primary intention is to influence people into adopting a certain model of behavior by convincing the addressees of the "inner goodness"[11] of their model.

Lastly, particularly in Western legal systems, *legal actors* can be defined as institutional actors primarily aiming at influencing the legal system, and therefore mainly focusing on the latter's logical structure. Like for political actors, the main goal of legal actors is to exercise power, that is, to force people to do things that they otherwise are not willing to do. As pointed out by Hans Kelsen, both law and politics try to make people do something, the law being "a social order, that is to say an order regulating the mutual behavior of human beings".[12] However, legal actors, when dealing with a statute or legal precedent, consider these normative documents as exercising their (binding) power on the addressees only as (and as long as they are) part of a larger hierarchical system of norms with a similar (legal)

[10] Alec Stone Sweet, 'Constitutional Courts' in Michel Rosenfeld and András Sajó (eds), *The Oxford Handbook of Comparative Constitutional Law* (OUP 2012) 823; and Wojciech Sadurski, *Rights Before Courts: A Study of Constitutional Courts in Postcommunist States of Central and Eastern Europe* (Springer 2005) 65–74. See also Albert H Y Chen and Miguel Poiares Maduro, 'The Judiciary and Constitutional Review' in Mark Tushnet, Thomas Fleiner, and Cheryl Saunders (eds), *Routledge Handbook of Constitutional Law* (Routledge 2015) 97–98 (as to the distinction between an American "decentralised" model and a continental European one of a more "centralised" character).

[11] Raz (n 4) 101; Niklas Luhmann, *Law as a Social System* (OUP 1994) 370; Gunther Teubner, 'Substantive and Reflexive Elements in Modern Law' (1983) 17 Law and Society Review 239, 259. See also Jürgen Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (The MIT Press 1998) 266. *cf* Frederick W Frey, 'The Problem of Actor Designation in Political Analysis' (1985) 17 Comparative Politics 127, 127–*12*9.

[12] Hans Kelsen, 'Law, State, and Justice in the Pure Theory of Law' in Hans Kelsen, *What is Justice? Justice, Law and Politics in the Mirror of Science* (University of California Press 1957) 289. See also Max Weber, *The Theory of Social and Economic Organization* (Free Press 1964) 152.

nature, and with specific (legal) rules to be used for interpretation, application, and creation (legal reasoning).[13]

III. The Relevance of Judicial Activism in Constitutional Courts Compared to Other Courts

The 'creative' interpretation of legal texts, or judicial activism, is a phenomenon generally characterizing all courts in a legal system, from the lowest county court to the highest court. However, judicial activism becomes a 'more evident' issue when addressed by Constitutional Courts (and the highest courts in general, e.g., the French Court of Cassation). First, the judicial activism performed by Constitutional Courts tends to become more visible as it often deals with fundamental questions relevant to all the other spheres of a society, that is, questions that by their nature can affect not only the legal or political arenas, but every arena and every person within a national community. Decision-making on the part of Constitutional Courts focuses on the fundamental laws of a community, and one of the features of a well-established democracy is the (actual or potential) intrusion of the law in all aspects of community life.[14] Moreover, the typically high degree of social legitimacy that Constitutional Courts are given by all actors within a certain community, contributes to rendering every decision taken by these courts more than simply a question of law and politics, but rather a question concerning the fundamental legal and political shape of that community.

When judicial activism takes place at the constitutional level, its consequences can reach the community at large, or at least – using Luhmann's terminology – it creates "noises" that can "disturb" the internal work (*autopoiesis*) of all the other subsystems.[15] This observation is valid for both Constitutional Courts with a jurisdiction of abstract review and those with a more concrete review jurisdiction. For example, in a specific euthanasia case submitted to its attention by a lower judge, a Constitutional Court's decision can face issues relevant for cultural or religious subsystems, namely the general question of how far the State can interfere with individual rights, or whether the 'right to live' encompasses also

[13] Habermas (n 11) 233–234; Kaarlo Tuori, *Critical Legal Positivism* (Ashgate Publishing 2002) 36–39; Luhmann (n 11) 188 (and his idea of "juridical rationality"). See also Max Weber, *Economy and Society: An Outline of Interpretive Sociology* (University of California Press 1978) 657.

[14] Ronald Dworkin, *Law's Empire* (Belknap Press 1986) vii; Marc Galanter, 'Law Abounding: Legalisation Around the North Atlantic' (1992) 55 Modern Law Review 1, 13–14. See also Lawrence M Friedman, *The Republic of Choice: Law, Authority, and Culture* (Harvard University Press 1998) 15.

[15] Luhmann (n 11) 80–84; Gunther Teubner, *Law as an Autopoietic System* (Blackwell 1993) 64–99. See also Matthew E K Hall, *The Nature of Supreme Court Power* (CUP 2010) 156–165.

the necessity for the public authorities to protect this right against the will of the 'owner' of such right.

The second factor underlying the general importance of judicial activism on the part of the Constitutional Courts relates to their being the highest judicial body of a legal system. It is true that judicial activism is important throughout all the levels of the judicial system. However, in the majority of cases, it is questionable whether lower level judicial law-making, though directly important for the community, can actually influence the higher levels.[16] For instance, a county court's law-making interpretation of a specific county council directive on shop licenses to allow for a certain kind of business in town tends to have an impact confined to the local implementing administrative offices, due to its limited jurisdiction. On the other hand, even such a simple issue can in the exceptional case have reverberations throughout the entire judicial system, for example, when it is a question of a pornography shop and freedom of speech.

When it comes to Constitutional Courts, their influence often transcends all the lower and intermediate levels, affecting the entirety of the legal structure, that is, including the lower structure, and a large part of the political world. This influence, and the resulting relevance of constitutional judicial activism over everything that is legal and political, is due to both the structure of the constitutional system itself (e.g, its being hierarchical) and the typically high degree of legal legitimacy that these types of courts have gained historically among the actors within the legal arena.[17] The centrality of the law-making of the Constitutional Courts is also recognizable in the fact that, when discussing judicial activism, most critical voices – whether in academia, politics, or the judiciary – end up having these Courts (not the lower ones) as their main targets.

IV. Converting the Hybrid Nature of Constitutional Courts Into a Theoretical Problem

When dealing with the issue of Constitutional Courts in relation to politics, the first question is where these institutional actors should be positioned on an imaginary map of two ideal-typical continents, namely the legal world and the

[16] Stone Sweet (n 7) 21–22. *cf* Lynn Mather, 'Theorizing about Trial Courts: Lawyers, Policymaking, and Tobacco Litigation' (1998) 23 Law & Social Inquiry 897; Jennifer Bowie and Elisha Carol Savchak, 'Understanding the Determinants of Opinion Language Borrowing in State Courts in the United States' in Susan M Sterett and Lee Demetrius Walker (eds), *Research Handbook on Law and Courts* (Edward Elgar Publishing 2019) 277.

[17] Habermas (n 11) 238–286; Marc Bühlmann and Ruth Kunz, 'Confidence in the Judiciary: Comparing the Independence and Legitimacy of Judicial Systems' (2011) 34 West European Politics 317, 332. See also Chen and Maduro (n 10) 100–101.

political world. This question is legitimate, since it is easy to see the importance of the role played by the courts in general and the Constitutional Courts in particular within both the legal and political systems.

As far as the legal system is concerned, Constitutional Courts are ranked at the top, being the supreme and ultimate interpreters of the constitution and consequently of the constitutionality of different law-making measures, in particular (but not exclusively) statutes. In other words, constitutional review is the legal competence allowing such courts to enjoy an exclusive decision-making power and a legal superiority in relation to the other branches of power, to an extent which is often unknown to most of the other judicial bodies within a national community. In addition to this lofty position within the legal structure, Constitutional Courts also tend to occupy a dominant place in the political building characterizing a democratic form of state. Constitutional Courts are entrusted by the political system (and, through it, by the community as such) to act as the ultimate guardians of the basic values that inspired the founding fathers and mothers when writing the fundamental documents (or in establishing the fundamental customs) underpinning and regulating the life of the political community.[18]

However, if the Constitutional Courts are seen from the perspective of the relations of law and politics, one can assert that they actually occupy a third position, at a much deeper level, functioning as a sort of transfer point between the legal and political worlds. If one considers the primary position occupied by Constitutional Courts (which is often implicit in the building of a modern democracy), this location can be identified as a bridge between values produced in the political world and in legal thinking. As stated by a political scientist,

> "[*c*]*onstitutional courts act systematically both in the legal and the political systems*. Almost every judgment has some consequences on the legal system (e.g., the abrogation of an unconstitutional law) and the political system (e.g., the retroactive defeat of the parliamentary majority that enacted this law)".[19]

As previously seen, the primary function of a Constitutional Court is constitutional review, that is, to continuously monitor the compatibility of

[18] Conrado Hübner Mendes, *Constitutional Courts and Deliberative Democracy* (OUP 2015) 73–82. See also Cass R Sunstein, *Legal Reasoning and Political Conflict* (OUP 1998) 82–83; Robert Alexy, 'The Dual Nature of Law' (2010) 23 Ratio Juris 167 (as to this dualistic position of the legal discourse in general).

[19] Michael Hein, 'Constitutional Conflicts between Politics and Law in Transition Societies: A Systems-Theoretical Approach' (2011) 3 Studies of Transition States and Societies 3, 17 *(emphasis in original)*. E.g., Martin Shapiro and Alex Sweet Stone, *On Law, Politics, and Judicialization* (OUP 2002) 81; Ralf Michaels, 'American Law (United States)' in Jan M Smits (ed), *Elgar Encyclopedia of Comparative Law* (2nd edn, Edward Elgar 2012) 75–76; or Nevil Johnson, 'The Interdependence of Law and Politics: Judges and the Constitution in Western Germany' (1982) 5 West European Politics 236, 239–244.

legislation and other normative measures with the basic values as announced in the constitution or other fundamental laws. While this role assigned to the Courts is quite undisputed, it carries an underlying problem of conflicting logic. At one end, the legal message of the Constitution, namely the models of behaviors prescribed, is heavily affected by the fact that constitutions are not only written by political actors (as are most legal measures), but are also often the product of extremely complex political compromises or very general political statements. Their being a political product, the constitution or fundamental law tends to be written less in legal terms, that is, as (at least in their intention) 'if x then y' or 'either/or' statements, and more as political messages, specifically in terms that resemble political propaganda, where the fundamental goals are models of behavior that the political actors want to be 'realised' in the community itself.[20] At the other end, constitutional documents are for historical reasons regarded as the highest sources of law in Western legal systems or, in other words – at least from a legal perspective – the constitution is legal in nature, that is, it is binding towards the addressees. As constitutions are designed as legal documents, they are treated as legal sources, having the strongest binding force on all the national law-making and law-applying agencies and on the community as well.[21] In other words, constitutions or fundamental laws in general have contents that tend to be dominated by the logics of the political discourse, but that are inserted into shells which have the form of laws and legal logics and shaping – somehow setting the agenda (at least as a direct effect) for the entire legal arena of a certain community.[22] As recently stated by a legal scholar,

> "[c]onstitutions structure the relationship of law and politics. They politicise the production of law, by connecting the legal system to a political process, and they legalize this political process through its obligation to superior legal rules".[23]

Since the primary goal of a Constitutional Court is to 'protect' such documents, it is easy to understand how this institutional actor tends to end up being a legal player, but inclined towards the political world. Three different

[20] E.g., Rudolf Streinz, 'The Role of the German Federal Constitutional Court: Law and Politics' (2014) 31 Ritsumeikan Law Review 95, 103–*104*; Jeremy Waldron, *Law and Disagreement* (OUP 2001) 220–222.

[21] E.g., András Sajó and Renáta Uitz, *The Constitution of Freedom: An Introduction to Legal Constitutionalism* (OUP 2017) 23–25; John O McGinnis and Michael B Rappaport, *Originalism and the Good Constitution* (Harvard University Press 2013) 130–*132*.

[22] Robert Alexy, *A Theory of Constitutional Rights* (OUP 2002) 349–*351*; and Martin Loughlin, 'Fundamental Law' in Miguel Nogueira de Brito and Luis Pedro Pereira Coutinho (eds), *The Political Dimension of Constitutional Law* (Springer 2020) 8–19.

[23] Christoph Möllers, 'Legality, Legitimacy, and Legitimation of the Federal Constitutional Court' in Matthias Jestaedt *and others* (eds), *The German Federal Constitutional Court: The Court Without Limits* (OUP 2020) 143.

features affect Constitutional Courts, making them actors that, though having their feet in the legal world, tend to lean heavily towards the political arena. First, from a legal theoretical perspective, Constitutional Courts are legal actors leaning into politics from an *institutional perspective*, that is, from the perspective of where these courts are positioned among the different organisations in a certain community which has as its primary goal governing the behavior of individuals, and are characterised as being permanent as well as making and enforcing rules governing human behavior. A Constitutional Court maintains its role as a legal institution, that is, an organisation constructed to safeguard certain important legal issues from a legal perspective, not the political opportunities that statutes create.[24]

At the same time, a Constitutional Court indirectly places the activities and operations of political actors, such as national or local assemblies, under scrutiny. It is true that its evaluation is directly legal in nature, but it is also true that the law is the main voice of political actors – at least in a democratic form of a state that has adopted the rule of law. Each time Constitutional Courts modify, approve, or even remain silent as to that which political actors have expressed through the law, the courts operate in the political institutional arena, particularly by allowing or disallowing certain political actors to produce statements that are directly relevant to and binding for the entire community from which such actors have (directly or indirectly) been chosen.[25] In other words, Constitutional Courts are legal institutional actors because of their being a court, but, at the same time, they are gatekeepers in relation to the political world, allowing the actors in the latter to be heard (or not) in the legal world.

Alongside this institutional factor, concerning the location of Constitutional Courts among the different actors, a second factor operates from a *structural perspective* in such a way as to render Constitutional Courts as legal actors heavily leaning into the political arenas. It has been mentioned how Constitutional Courts reside outside of the ordinary court system and are independent from other branches of the public authorities. However, Constitutional Courts always tend to present a certain 'structural cohesion' with the actors belonging to the political arena. This means that almost all Western legal systems have foreseen that political

[24] Dieter Grimm, 'What Exactly Is Political about Constitutional Adjudication?' in Christine Landfried (ed), *Judicial Power: How Constitutional Courts Affect Political Transformations* (CUP 2019) 310–311. *cf* Robert A Dahl, 'Decision Making in Democracy: The Supreme Court as National Policy-Maker?' (1957) 6 Journal of Public Law 279, 279; Kevin T McGuire, 'The Institutionalization of the U.S. Supreme Court' (2004) 12 Political Analysis 128, 129–135 (as to the traditional political science perspective on the issue).

[25] Grimm (n 24) 308–309; Nuno Garoupa and Tom Ginsburg, 'Building Reputation in Constitutional Courts: Political and Judicial Audiences' (2011) 28 Arizona Journal of International & Comparative Law 539, 541.

actors, either as legislators or within the executive branch, can have partial (as in Italy) or total (as in the US) control as far as concerns the individuals who are to sit as justices in the Constitutional Courts. As a consequence, the political arena and the ideologies prevailing within it affect and to a certain extent overlap with the structure of the courts and their fundamental components, by means of the legal power to decide who will be justices.[26]

Despite this important political influence in deciding the structure of Constitutional Courts, these courts cannot be considered as having a purely and exclusively 'political structure'. Though the justices sitting in such courts can (and often are) politicised individuals, they nevertheless come from the legal world, that is, the predominant feature of individuals sitting in Constitutional Courts is normally that they are chosen among lawyers or individuals with formal education in law. In other words, even when all justices are chosen based on political considerations and personal political ideologies and affiliations, the selection process is limited (either by law or by constitutional customs) to individuals trained at least formally in the art of law, for example, holding a law degree. Moreover, most (but not all) of the time, the recruitment procedures require that the candidates have some experience from the judiciary branch at a high level.[27]

Lastly, Constitutional Courts can be seen as legal actors inclined towards the political world from a *functional perspective*, that is, by observing the function these courts have in the relations between lawyers and politicians.[28] Viewed from this functional perspective, one can note how Constitutional Courts perform an intermediary function between these two arenas. As briefly sketched above, one of the major contributions of a Constitutional Court to its community is mediating between the highly political statements present in the constitution. The articles of a constitution tend to be dominated, from a legal perspective, by the rationality of

[26] Carlo Guarnieri and Patrizia Pederzoli, *The Power of Judges: A Comparative Study of Courts and Democracy* (OUP 2003) 138–141; Walter F Murphy and Charles Herman Pritchett, *Courts, Judges, and Politics: An Introduction to the Judicial Process* (4th edn, Random House 1986) 139; Nicola Ch Corkin, *Europeanization of Judicial Review* (Routledge 2014) 115–116. E.g., Donald P Kommers, 'American Courts and Democracy: A Comparative Perspective' in Kermit L Hall and Kevit T McGuire (eds), *Institutions of American Democracy: The Judicial Branch* (OUP 2006) 207.

[27] Stone Sweet (n 7) 45–49; Klaus Stüwe, 'The U.S. Supreme Court and the German Federal Constitutional Court: Selection, Nomination, and Election of Justices' in Rogowski and Gawron (n *9*) 242–244. cf Peter H Russell, 'Conclusion: Judicial Independence in Comparative Perspective' in Peter H Russell and David M O'Brien (eds), *Judicial Independence in the Age of Democracy: Critical Perspectives from Around the World* (University of Virginia Press 2001) 303–304.

[28] In this article, "function" refers to the "function as effects" (as different from "function as purpose") of a certain institution on a certain environment, or, as in this case, the concrete outcomes that the work of Constitutional Courts have on both legal and political structures. Brian Z Tamanaha, *Law as a Means to an End: Threat to the Rule of Law* (CUP 2006) 245–249.

the political discourse; at the same time, they are 'legally relevant' concepts and categories, that is, concepts binding public officials and the community in general through their observance of the parameters of the rationality required by a legal system.[29]

This mediating role played by Constitutional Courts is not only directed at the legal world, where the Courts define for its actors in legal terms what the general statements of goals in the constitutional documents or practices mean, for example, by guiding a justice in the interpretation of constitutionally questionable statutes. The mediating function is also aimed at the political arena, as the decisions of Constitutional Courts set the legal frameworks that the political actors ought to respect in their law-making.[30]

Since a constitution is the product of the will of a community (through their political representatives), at least in theory, a Constitutional Court in a democratic state has the function of mediating to the community and, in particular, to its political representatives, the value message that this same community and its political actors originally adopted, but now in terms of a legal message, that is, a message primarily directed at the actors given the duty of implementing the legal rules as interpreted (or accepted) by the Constitutional Court. As in particular the American legal realists and Alf Ross have pointed out, judges in general play a decisive role as the point of passage where the "law-in-books" becomes the "law-in-action," that is, the normative apparatus of rules felt as binding by the population or by the public officers.[31]

In this case, Constitutional Courts have the primary function of translating into binding norms for the political actors and the community, the law-in-books that the political actors have enacted. Constitutions tend to become documents where the political origins of the law, a typical feature of contemporary law, surface more clearly than in other legal documents (e.g., a statute regulating taxation law). Constitutions are often used not only as a legal document grounding a new legal

[29] Joseph Raz, 'On the Authority and Interpretation of Constitutions: Some Preliminaries' in Raz (n 4) 369–370; and Jiří Přibáň, *Legal Symbolism: On Law, Time and European Identity* (Routledge 2007) 21. E.g., Dworkin (n 14) 370–371.

[30] Torbjörn Vallinder, 'When the Courts Go Marching In' in C Neal Tate and Torbjörn Vallinder (eds), *The Global Expansion of Judicial Power* (New York University Press 1995) 13–24. E.g., Ulrich K Preuß, 'Die Wahl der Mitglieder des BVerfG als verfassungsrechtliches und -politisches Problem' (1988) 19 Zeitschrift für Parlamentsfragen 389, 389–91; Mark Tushnet, *Taking the Constitution Away from the Courts* (Princeton University Press 1999) 7–9.

[31] Karl N Llewellyn, *Bramble Bush: On Our Law and Its Study* (Oceana Publications 1951) 150; Roscoe Pound, 'Law in Books and Law in Action' (1910) 44 American Law Review 12, 35–36; Alf Ross, *On Law and Justice* (University of California Press 1959) 75–77. See also Ronald Dworkin, *Freedom's Law: The Moral Reading of the American Constitution* (Harvard University Press 1996) 7–15; and Waldron (n 20) 262.

system, but also as a primary form of "political symbol", that is, a message to the community from the political actors as to which fundamental values the state/community is based upon.[32] Moreover, and as a consequence of this partially political nature, legal language in the Constitution tends to be interspersed with political language.[33] A classic example of this is the article of the Italian constitution stating that property ownership is guaranteed by the law as long as it fulfills its social function.[34]

This being the situation, where judges are the intermediaries between the "paper law" and "real law," with the "paper law" being the Constitution – a mixture of political statements and legal concepts – it is not surprising that Constitutional Courts, more than other branches of the judiciary, become the law-making actors by being the interpreters of the law. As already pointed out by many legal scholars, the interpretation of the law that is typical for a court in general can become a law-making power.[35] In the case of the Constitutional Courts, this phenomenon is more evident, because the legal text against which the interpretation of the statutes has to take place (namely the constitutional document) is so vague that the clarification of its content and of its borders becomes law-making (at least if seen from a legal perspective) directly applicable in concrete cases (as regards the concrete constitutional review) or in general (as regards the abstract constitutional review).[36] In summary, Constitutional Courts, for all the reasons mentioned above, are a special type of institutional actors which, though positioning themselves among the legal actors, that is the actors aiming at interpreting and applying the law, tend to lean heavily into the political world of law-making, given that

[32] Raz (n 29) 343–344.

[33] ibid 365–366; Habermas (n 11) 388–389. See also Mark Tushnet, 'Abolishing Judicial Review' (2011) 27 Constitutional Commentary 581, 585–586.

[34] Constitution of the Italian Republic, Article 42: "Private property is recognised and guaranteed by the law, which prescribes … its limitations so as to ensure its social function and make it accessible to all". See also Giuseppe Portonera, 'The Problem of Squatting in Italy: A New Approach by the Courts' (International Index of Property Rights, 2019) 4–5 <https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3472293_code3097842.pdf?abstractid=3472293&mirid=1> accessed 14 March 2021.

[35] E.g., Frederick Schauer, 'Opinions as Rules' (1986) 53 University of Chicago Law Review 682, 684; Hart, *The Concept of Law* (n 6) 132–136; Neil MacCormick, *H.L.A.* Hart (2nd edn, Stanford University Press 2008) 157–159. See also Aharon Barak, 'A Judge on Judging: The Role of a Supreme Court in a Democracy' (2002) 116 Harvard Law Review 19, 62; John Hart Ely, *Democracy and Distrust: A Theory of Judicial Review* (CUP 1980) 1–9.

[36] Gunther Teubner, 'And God Laughed ... Indeterminacy, Self-Reference and Paradox in Law' in Jean Pierre Dupuy and Gunther Teubner (eds), *Paradoxes of Self-Reference in the Humanities, Law and the Social Science* (Anma Libri 1991) 31. See also Stone Sweet (n 10) 827–828; Anna Gamper, 'Constitutional Courts and Judicial Law-Making: Why Democratic Legitimacy Matters' (2015) 4 Cambridge Journal of International and Comparative Law 423, 424–434.

Constitutional Courts through interpretation shape the legal panorama regulating a certain community.[37]

V. Relevance of The 'Political vs Legal Nature' Discussion as to the Constitutional Courts

Shifting attention to the topic central to this work, namely whether Constitutional Courts should be defined as primarily legal or political actors, a first reaction could be to question the importance of this issue. It appears to be a purely terminological matter, as the competence and jurisdiction accorded to Constitutional Courts, at least in well-established Western democracies, tend to be the same from a legal perspective, regardless of whether they are considered more political or more legal actors operating inside a certain system of powers. Whether they are seen primarily as legal or political actors, justices sitting on the highest benches will always be in charge of determining the constitutionality of statutes and, by doing this, will always be influenced by the political environment and prevailing political ideologies.[38]

However, the question presented here is not simply a definitional or academic problem. As often happens in legal matters, defining something or someone means attributing it with certain legal areas of competence and jurisdiction and, at the same time, limiting its capacity to operate in other legal areas. In other words, when it comes to legal issues, the classification of either a problem or a subject-matter means shaping it and, at the same time, restricting it.[39]

If one considers in particular Constitutional Courts and the definition of their nature as actors working in a certain environment, it has previously been seen that among their central tasks is 'checking' that the transformations of ideologies or values into law are done in accordance with (or at least not in gross contradiction of) the basic and often politically formulated principles enumerated in the constitution or fundamental laws of a certain community. The characterisation of Constitutional Courts as being either legal or political actors brings with it the identification of fundamental criteria, or in Max Weber's terminology, 'rationalities', that ought to govern this control over the constitutionality of the law-making that takes

[37] Luhmann (n 11) 235.

[38] Hans Kelsen, *The Pure Theory of Law* (University of California Press 1970) 3–5; Waldron (n 20) 144. E.g., Alexy (n 22) 366; Lawrence Baum, *American Courts: Process and Policy* (7th edn, Wadsworth 2013) 270–272.

[39] Timothy A O Endicott, 'Law and Language' in Jules L Coleman and Scott Shapiro (eds), *Handbook of Jurisprudence and Legal Philosophy* (OUP 2002) 935–968. See also Waldron (n 20) 229; Peter Goodrich, *Legal Discourse: Studies in Linguistics, Rhetoric, and Legal Analysis* (Palgrave Macmillan 1988) 2–3. E.g., Linda L Berger, 'Applying the New Rhetoric to Legal Discourse: The Ebb and Flow of Reader and Writer, Text and Context' (1999) 49 Journal of Legal Education 155, 155.

place in a certain legal system.[40] By normatively defining the nature and function of Constitutional Courts (e.g., what they 'ought' to be and to do), it becomes possible to answer the following normative question that is fundamental for every democratic legal system: What are the fundamental criteria that ought to guide a Constitutional Court when performing its task of constitutional review?

Considering that Constitutional Courts operate between the legal and political worlds, it is possible to identify two fundamental criteria, or rationalities, which could shape Constitutional Courts in their work. First, at least when seen from a legal perspective, Constitutional Courts have the option to primarily embrace a Weberian substantive rationality to resolve issues of constitutionality.[41] This choice would mean that, to reach the 'best' solution, justices would regard the legal system as primarily instrumental to the fulfillment of certain goals external to the system itself. In other words, Constitutional Courts ought to be ready to 'sacrifice' the internal rationality and rules traditionally superseding Western legal systems and reasoning, if and as long as this capitulation is directly functional to achieving the political, social, and economic values the courts intend, on various grounds, to insert into a certain community.

However, there is another possible ideal-type rationality or criterion that may guide Constitutional Courts in their work. As pointed out by Weber, in modern capitalist societies, the fundamental criterion inspiring the work of legal actors is formal rationality: they reach a decision or a legal solution based on the logical criteria internal to the legal system and with the purpose of maintaining its consistency, regardless of the actual effects in the surrounding environments.[42] This respect for formal rationality (or 'legality') exists and ought to exist, because, as Weber stated, it is directly functional and fundamental for legal actors (and judges in particular) to gain and maintain their legitimacy, that is, a high degree of probability that their decisions will be observed by the majority of addressees because they are considered 'correct' and, therefore, binding.[43]

The characterisation of a certain actor as legal or political is then always fundamental, at least from a legal perspective, to attach to a certain actor

[40] Weber (n 13) 650–658; Max Weber, *Max Weber on Law in Economy and Society* (Max Rheinstein tr, Simon & Schuster 1954) 63–64.

[41] Weber (n 13) 656–658. See also Reza Banakar, *Normativity in Legal Sociology:* Methodological Reflections on Law and Regulation in Late Modernity (Springer 2015) 219.

[42] Weber (n 13) 655. See also David M Trubek, 'Max Weber on Law and the Rise of Capitalism' (1972) 3 Wisconsin Law Review 720, 733; Anthony T Kronman, *Max Weber* (Stanford University Press 1983) 90.

[43] Weber (n 13) 654–658. See also Sally Ewing, 'Formal Justice and the Spirit of Capitalism: Max Weber's Sociology of Law' (1987) 21 Law & Society Review 487, 497–502. cf Arthur L Stinchcombe, 'Reason and Rationality' (1986) 4 Sociological Theory 151 (as to the substantive rationality being a *conditio sine qua non* for every formal rationality).

a certain criterion (or type of rationality, as in this article) that should guide it in its operations. This definition, however, is even more important in the case of Constitutional Courts, due to the position such courts occupy in modern democratic forms of political organisation. Constitutional Courts are certainly not the only actors whose nature can be and is widely disputed. For example, the legal nature of in-house attorneys is often heavily questioned; they are sometimes treated as simply facilitating a legal cover-up of economic and political programs.[44] However, the theoretical issue of normatively defining Constitutional Courts is fundamental: the decisions of these courts (and consequently the criteria inspiring them) are those that can shape the fundamental legal, but also political and social, features of an entire community, sometimes even more than the decisions made in the democratically elected assembly. For example, in deciding *Brown v. Board of Education* (1954), the Supreme Court of the United States shaped (at least as much as Congress did ten years later with the 1964 Civil Rights Act) the future of an entire national community as far as concerned unlawful structural discrimination based on ethnicity.[45]

It is also true – using Ronald Dworkin's famous metaphor – that Constitutional Courts write just one chapter in the chain novel that constitutes the valid law because after their decisions, their words will be interpreted by all the other actors, for example, legal scholars, lower judges, and law-makers.[46] However, even if the subsequent actors write a 'different' continuation of the novel, it is the Constitutional Courts that have the privilege of setting the agenda for future discussion.[47] For example, with *Brown v. Board of Education,* the US Supreme Court definitively opened the door to de-segregation, that is, it gave a strong push to put

[44] E.g., Robert L Nelson and Laura Beth Nielsen, 'Cops, Counsel, and Entrepreneurs: Constructing the Role of Inside Counsel in Large Corporations' (200) 34 Law & Society Review 457, 464–468; Prashant Dubey and Eva Kripalani, *The Generalist Counsel: How Leading General Counsel are Shaping Tomorrow's Companies* (OUP 2013) 66–67.

[45] *Brown v Board of Education of Topeka* 347 US 483 (1954); Michael J Klarman, *Brown v.* Board of Education and the Civil Rights Movement: The Supreme Court and the Struggle for Racial Equality (OUP 2007) 302–338. As to a more European context, see also Guarnieri and Pederzoli (n 26) 1–4; Stone Sweet (n 7) 66–70.

[46] Dworkin (n 14) 228–238. See also Mark J Richards and Herbert M Kritzer, 'Jurisprudential Regimes in Supreme Court Decision Making' (2002) 96 The American Political Science Review 305, 306 ("[L]aw at the Supreme Court level is to be found in the structures the justices create to guide future decision making: their own, that of lower courts, and that of non-judicial political actors"); Martin Shapiro, *The Supreme Court and Administrative Agencies* (Free Press 1968) 39.

[47] Luhmann (n 11) 406 (indicating how the main feature of modern constitutions is their "openness to the future"). See also Raz (n 29) 338–343.

into the trash can all the attempts to retain in American society the racist principle 'separate but equal'.

Many other aspects, of both political and legal character, underscore the necessity of coming forward with a clear definition of what kind of actors Constitutional Courts should be. From a political perspective, the definition of a Constitutional Court is important as it clarifies, and therefore, partially prevents, possible points of collision between the highest powers in a community. Pointing out the basic features and criteria that should inspire the work taking place in the courts allows for a better and more precise control of the activity of the courts by political authorities, for example, in the form of offering a clear matrix to parliamentary committees or investigators against which to evaluate certain constitutional judicial decisions. In other words, this legal theoretical definition more clearly pinpoints a fundamental actor on the political map, either as primarily promoting certain ideologies that vary over time (if defined as a political actor) or as primarily attempting to maintain one single and established legal ideology, namely the rule of law (if defined as a legal actor).

Characterizing the nature of Constitutional Courts is also important from a legal perspective, as this allows normative fixing of what type of rationality Constitutional Courts ought to apply in their work. At one end, justices sitting on the highest benches may be defined primarily as legal actors. Accordingly, from a legal perspective, the legality of their decisions in "hard cases" can and should be questioned, even by lower courts, when their legal reasoning is mainly grounded on the goal of implementing values they consider as immanent in a community, although such values do not explicitly appear in the constitutional documents or fundamental laws. This critique of the legality of their decisions can and ought to be performed, in particular when the realisation of values is done at the expense of the traditional criteria superseding the legal reasoning (for example, consistency or respect for previous decisions on similar matters), that is, the only type of reasoning on which legal actors in modern democracy have a legitimate domain. For example, if a court decides within a quite narrow timeframe in diametrically opposite directions in similar cases or issues, it can be directly criticised from a legal perspective for violating a fundamental principle of Western legal systems, namely equal treatment of individuals under identical circumstances.

At the other end, in the event that Constitutional Courts are defined primarily as political actors, the possibility of holding them responsible from a legal perspective for doing something 'illegal' is more restricted. If they are considered political actors, it is not possible to 'force' the courts to decide in accordance, or at least in consistency, with previous decisions, although it is always possible to legally criticize Constitutional Courts for violating certain basic rights guaranteed

in the constitution. One privilege accorded to political actors in general is that they can change their value system without being held responsible (at least legally) for this. If a political party or national assembly decides to pursue values other than those previously planned, it cannot be criticised or held responsible from a legal perspective.[48]

Lastly, another element highlights the importance of normatively setting the nature of Constitutional Courts in Western legal systems. These courts symbolize (and stretch to the limits) an underlying feature typical of most legal actors operating in contemporary Western legal systems: their position in between the political world, where values (or models of society) are created, and the legal world, through which those values have to pass in order to be implemented into a community.

Each of the individuals forming the skeletal structure of the legal actors is educated in the law and such an education is almost always a formal requirement to becoming a part of that group of actors. The individuals composing the legal actors are, in other words, all educated in the idea that law, although highly politicised, has certain features that distinguish it from purely political statements.[49] For justices working in Constitutional Courts, it is the same as for most legal actors: they operate within the legal system, but with the knowledge that law is instrumental to introduce into a community models of behaviors or values embraced by their political source (e.g., legal experts working in political parties) or their economic source (e.g., in-house attorneys for large corporations). This feature of the law in the Western legal systems, that is, it always being functional to something else, then forces legal actors in general to always take into consideration the value systems (and the underpinning ways of reasoning) affecting the origins, development, and environment in which the making or application of the law is taking place.[50]

In short, the importance of defining Constitutional Courts as either legal or political actors lies also in the fact that such institutional actors represent,

[48] Stone Sweet (n 7) 62. See also John Ferejohn and Pasquale Pasquino, 'Constitutional Courts as Deliberative Institutions: Towards an Institutional Theory of Constitutional Justice' in Wojciech Sadurski (ed), *Constitutional Justice, East and West: Democratic Legitimacy and Constitutional Courts in Post-Communist Europe in a Comparative Perspective* (Kluwer Law International 2002) 23.

[49] Tuori (n 13) 161; Roger Cotterrell, *Law's Community: Legal Theory in Sociological Perspective* (Clarendon Press 1995) 108–110. See also Teubner (n 15) 33; Andrew D Abbot, *The System of Professions: An Essay on the Division of Expert Labor* (University of Chicago Press 1988) 52–57; Neil MacCormick, *Legal Reasoning and Legal Theory* (Clarendon Press 1978) 188.

[50] Cotterrell (n 49) 281–284; Richard Posner, 'The Decline of Law as an Autonomous Discipline: 1962–1987' (1987) 100 Harvard Law Review 761, 773–774; Francis J Mootz III, '"Die Sache": The Foundationless Ground of Legal Meaning' in Jan M Broekman and Francis J Mootz III (eds), *The Semiotics of Law in Legal Education* (Springer 2011) 48–49. See also Raz (n 4) 99–100; Jeremy Waldron, 'Legislation, Authority, and Voting' (1996) 84 Georgetown Law Journal 2185, 2198.

better than many others, the difficult situation in which lawyers in general operate nowadays. While they are educated in the law and employed to build, interpret, and apply the law, legal actors operate under a constant and extreme pressure that pushes them towards a disregard for what are considered the characterizing elements of a Western or Western-like legal system (predictability, certainty, rule of law, and so on), in order to instead fulfill political (or other non-legal) goals.[51]

VI. A Possible Legal Theoretical Solution

Part III ('Converting the Hybrid Nature of Constitutional Courts into a Theoretical Problem') showed how Constitutional Courts can be considered as actually belonging to the legal world, but strongly leaning towards the political arena. Part IV ('Relevance of the 'Political vs. Legal Nature' Discussion as to the Constitutional Courts') pointed out that it is important, for several reasons, to normatively 'insert' the Constitutional Courts into either of these two ideal-typical boxes, that is, to establish which of the two natures (political vs. legal) should dominate their work and should be used as a starting point for investigating and (if warranted) criticising their decision-making.

A possible perspective from whence to begin the journey to answer this fundamental question is legal theory. Due to the central position and function that Constitutional Courts play in contemporary legal systems, legal theory has devoted many and important writings to this topic: most contemporary legal theories have discussed the issue of what Constitutional Courts are, somehow being forced to tackle this question due to the impact of these courts' decisions on the law and society at large.[52]

Though the journey will continue along legal theoretical paths, it is helpful here to take a slight detour to a sociological distinction between the *institutional* (or *organisational*) *position* (or *status*) of a certain actor (or agent) and the *function-effects* of

[51] Gunther Teubner, 'The Transformation of Law in the Welfare State' in Gunther Teubner (ed), *Dilemmas of Law in the Welfare State* (Walter de Gruyter 1986) 6–7; Lawrence Friedman, 'Introduction' (2003) 4 Theoretical Inquiries in Law 437, 446; Kaarlo Tuori, 'Legislation between Politics and Law' in Luc J Wintgens (ed), *Legisprudence: a New Theoretical Approach to Legislation* (Hart Publishing 2002) 100–101.

[52] E.g., the debate between Robert Alexy, *The Argument from Injustice: A Reply to Legal Positivism* (Clarendon Press 1992) 5–7, and John Finnis, 'Law as Fact and as Reason for Action: A Response to Robert Alexy on Law's "Ideal Dimension"' (2014) 59 The American Journal of Jurisprudence 85, 105–106 (as to the decision by the German Federal Constitutional Court on citizenship *BVerfGE 23 – Ausbürgerung I* [1968], 98, 98–113); or Dworkin (n 14) 387–392 and Duncan Kennedy, *A Critique of Adjudication (fin de siècle)* (Harvard University Press 1998) 127–128 (as to the US Supreme Court's ruling in *Brown v Board of Education*).

that actor's work.[53] The institutional position of a certain actor intends to signify the position occupied by a certain actor operating inside a larger environment (or organisational structure).[54] This positioning, as far as concerns judicial bodies, is mainly a combination of the operation of two (often overlapping) factors: the degree of legitimacy that judicial bodies enjoy, indicating where along the spectrum of power judges are inserted (vertical positioning), and the distribution of power as sanctioned in the law, which indicates where, at the stage assigned by the legitimacy, the judicial body is located (horizontal positioning). The function-effects of an actor's work simply refers to the impact that the work of the actor has on the environment.[55] These effects can be of different ideal-typical natures. They can be *intended*, where they correspond to the original goal that the actor had in mind when starting the work, or *unintended*, where they do not (fully or partially) correspond to the original motive of the action.[56] Effects can also be in the form of either *outputs* or *outcomes*.[57] Outputs are the impacts (intended or unintended) a certain action has inside the ideal-typical arena in which the action has taken place (e.g., the effect of a court decision on the legal right of the convicted party to appeal). Outcomes, by contrast, mark the effects (intended or unintended) such

[53] Talcott Parsons, *The Social System* (Free Press 1951) 25. As to an application of this distinction within the legal discourse (specifically to the judiciary), see also Robert F Williams, 'In the Supreme Court's Shadow: Legitimacy of State Rejection of Supreme Court Reasoning and Result' (1984) 35 South Carolina Law Review 353, 397–402.

[54] Philip Selznick, *Leadership in Administration: A Sociological Interpretation* (University of California Press 1957) 17–22. *cf* Ota Weinberger, *Law, Institution and Legal Politics: Fundamental Problems of Legal Theory and Social Philosophy* (Springer 1991) 18–24; Neil MacCormick, 'Norms, Institutions, and Institutional Facts' (1998) 17 Law and Philosophy 301, 324–331 (where "institutions" within a traditional legal context instead refers to the regulatory tools, e.g., contract or ownership).

[55] Roger Cotterrell, *The Sociology of Law: An Introduction* (2nd edn, Butterworths 1992) 72–73.

[56] Robert K Merton, 'The Unanticipated Consequences of Purposive Social Action' (1936) 1 American Sociological Review 894.

[57] This separation of outputs from outcomes is an adaptation of the results reached by a long series of studies developed in political science. David Easton, *A Systems Analysis of Political Life* (Wiley 1965) 351–352; Gabriel Abraham Almond, G Bingham Powell, and Robert J Mundt, *Comparative Politics: A Theoretical Framework* (Harper Collins 1993) 20; Jan-Erik Lane and Svante T Ersson, *The New Institutional Politics: Outcomes and Consequences* (Routledge 2000) 60–62. E.g., Selden Biggs and Lelia B Helms, *The Practice of American Public Policymaking* (Routledge 2015) 370–371.

impacts have on the surrounding environment (e.g., the effect of a court decision on the economic situation of the convicted party's family).[58]

If one considers Constitutional Courts in light of these distinctions between institutional positions and function-effects (and the different types of effects), one can see that the dominant features of the courts are of a legal nature. Starting from the *institutional position*, Constitutional Courts are, first and foremost, 'courts'. This label means that their rulings are considered binding by the vast majority of the addressees, not because of the content of the decisions, that is, the models of behaviors they aim to impose on a community, but because they are legal normative decisions. This means that they are rulings that ought to be obeyed because they are produced by a legally formed body, which is entrusted, in the forms prescribed by the law, with the legal power to produce such binding decisions. In contrast to political actors, such as political parties or lobby groups, the consideration and respect for the work of Constitutional Courts is not mainly based on the intrinsic values promoted by its decisions, such as the 'popularity' of a certain political program. The respect, or legitimacy, is given to the decisions due to the legal form such rulings take, and the forms that have been observed while producing the decisions and choosing the individuals (e.g., judges) in charge of making such decisions. In other words, Constitutional Courts keep their position and 'job' in the community as the highest dispute-resolving actor as long as they are able to

[58] As to the distinction between "consequences" and "juridical consequences", Neil MacCormick, 'On Legal Decisions and Their Consequences: From Dewey to Dworkin' (1983) 58 New York University Law Review 239, 247–251. E.g., Tomas M Koontz and Craig W Thomas, 'Measuring the Performance of Public-Private Partnerships: A Systematic Method for Distinguishing Outputs from Outcomes' (2012) 35 Public Performance & Management Review 769, 772 (Figure 1). In reality, these different ideal-types (intended outputs, unintended outputs, intended outcomes, unintended outcomes) almost always tend to be mixed with each other, e.g., in the form of court decisions that have intended and unintended outputs and outcomes simultaneously. Despite this overlapping in the real world, such ideal-types can be useful analytical tools in order to reveal specific tendencies of an actor to operate in one environment instead of another in order to gain certain effects, and then normatively choosing the type of rationality more suitable for that purpose, e.g., indicating the line that the actors, as belonging to a certain arena, ought to pursue.

maintain their legal legitimacy, that is, a legitimacy gained in Western legal systems mostly by observing the paradigms of formal legal rationality (or 'legality').[59]

Naturally, this pushing of the Constitutional Courts into the 'legal box' cannot ignore the fact that justices have political sympathies. However, even when justices are strongly politicised, they still ought to operate with an eye to (and being forced to comply with the barriers of and limits as set by) the legal system. In other words, despite being connected to the political world, the judges sitting in the Constitutional Courts ought to observe the principles or paradigms established by the dominant legal culture (e.g., rule of law, bill of rights, separation of powers, due process and so on) in order not to lose their legitimacy among the addressees (and the community at large).[60]

As to their *function*, if one starts by considering the *intended* and *unintended outputs* of a certain decision by a Constitutional Court, the primary arena of operation of Constitutional Courts is the legal one. A Constitutional Court, by definition, evaluates legal issues, in particular the potential unconstitutionality of statutes or acts from law-making agencies. The outputs of the courts' deliberation are to decide whether certain legal rules of a lower dignity can still be considered as 'binding and therefore existing' legal rules. In particular, Constitutional Courts 'prove' the existence of these rules by evaluating whether they are compatible with the fundamental rules and principles enumerated in (or somehow derived from) constitutions and fundamental laws. This test, as pointed out before, is a typically legal problem since it can operate if, and only as far as, one axiomatically accepts the existence of an ascending structure of rules, where the lower rules, in order to

[59] Ely (n 35) ch 4; Habermas (n 11) 278–279. See also Hein (n 19) 17 ("Constitutional courts, *per se*, have some leeway for making decisions based on political criteria. However, if this margin is too wide, and if the court is dependent on the political interests of other state powers, constitutional conflicts will be provoked"); and Hans Kelsen, *General Theory of Law and State* (Routledge 2005) 117. cf Sadurski (n 10) 53–61 (pointing out, in order to be transformed into legitimacy, the necessity of formal rationality, or "legality" in his words, to be finalised to the realisation of certain values external to the legal discourse, e.g., substantive rationality).

[60] Niklas Luhmann, 'Operational Closure and Structural Coupling: The Differentiation of the Legal System' (1991) 13 Cardozo Law Review 1419, 1435. E.g., Alexy (n 22) 367. See also Ferreres Comella (n 9) 19 ("We cannot automatically claim that if a given institution strikes down statutes, it is really a 'legislative' body, whereas if it merely sets them aside for purposes of resolving disputes, it acts as a real 'court.' What matters is the sort of grounds -political or legal- on which the institution rests its decisions").

exist as legal (and therefore binding) rules, cannot be in conflict with the higher ones.[61]

As pointed out by Kelsen and other legal scholars, this presupposition is typical of the legal arena.[62] By contrast, the hierarchical structure in the political arena, though present (e.g., basic values vs. tactical choices), is not fundamental to give 'validity' to the lower types of decision. Tactical decisions taken by a congressional party are still considered 'valid' for the political line of a certain party (e.g., because such tactical decisions can strengthen the party's positions in an upcoming election), even if they are contrary to the fundamental values contained in the party program.

In contrast to political actors, Constitutional Courts are not totally free from what can be defined as the external borders of legal reasoning. 'External borders' are identified in particular as the no-cross limits of the legal culture of a certain community, limits which have to exist in order for the legal system as such to survive.[63] In a democratic free-market regime, for example, these no-cross borders can be defined as the fundamental legal principles (e.g., protection of private property) expressing the bedrock of political, cultural, and economic forces upon which the regime itself is created and to which it is functional.[64]

Political actors do not necessarily have to respect such external borders of legal reasoning. Actually, for many political parties, the primary and fully politically legitimised goal for their existence is to change or shift such external borders, for example, by eliminating the legal protection accorded to private property. The situation changes if one moves the focus to the *outcomes* of the decisions taken by Constitutional Courts, that is, the effects (intended or unintended) that their decisions have on the (non-legal) environments surrounding the (legal) one in which the courts operate. It is easy to see how the legal feature characterizing the function played by the Constitutional Courts here tends to disappear. Decisions

[61] Hart, *The Concept of Law* (n 6) 100–110; Hans Kelsen, *Introduction to the Problems of Legal Theory* (Clarendon Press 1997) 11 ("To comprehend something legally can only be to comprehend it as law"). See also Luhmann (n 11) 406–407.

[62] Kelsen (n 59) 110–113. See also Kelsen (n 38) 3–4 and 19; Robert S Summers, *Form and Function in a Legal System: A General Study* (CUP 2006) 313; and Neil MacCormick, 'Natural Law Reconsidered' (1981) 1 O*xford Journal of Legal Studies* 99, 108.

[63] Hart, *The Concept of Law* (n 6) 193–200; HLA Hart, 'Problems of the Philosophy of Law' in HLA Hart, *Essays in Jurisprudence and Philosophy* (Clarendon Press 1983) 112 (as to his idea of "minimum content of natural law"). See also Tuori (n 13) 177–183 (as to his idea of the "general legal principles" of the legal culture).

[64] E.g., Jeremy Waldron, *The Rule of Law and the Measure of Property* (CUP 2012) 103; Neil MacCormick, 'MacCormick on MacCormick' in Augustín José Menéndez and John Erik Fossum (eds), *Law and Democracy in Neil MacCormick's Legal and Political Theory:* The Post-Sovereign Constellation (Springer 2011) 23.

by Constitutional Courts almost always have effects outside the legal world, for example, in the cultural, economic, or political spheres.[65] In other words, as far as concerns the outcomes of their decisions, Constitutional Courts have certain similarities with political actors such as the government or national assemblies. What is more, Constitutional Courts ultimately attempt with their decisions (consciously or unconsciously) to impose certain models of behaviors or values upon a community.

Despite this leaning into the political arena, Constitutional Courts should be considered to have a legal nature, that is, as legal actors (and act as though they are). First, the grouping of an actor under a certain terminological roof has to be done primarily based on its *institutional location* and the function-effects of its work, in particular the *intended outputs* that its actions produce.[66] If one were to look at the outcomes, the analytical possibility of grouping actors in ideal-typical arenas, and the resulting possibility of identifying some normative criteria upon which to evaluate and criticize their work, would disappear. Outcomes of decisions almost always tend to spread in different directions and, especially for unintended outcomes, it is often not possible to determine in which area a certain action has had its major impact, particularly after a long period of time. For example, a decision made by a large corporation can have relevant outcomes in the religious or cultural fields, but it would be quite strange to define such corporations as primarily religious or cultural actors, and consequently impose upon the CEO or the board of directors religious or cultural criteria according to which to evaluate their work.

It goes without saying that the positioning of Constitutional Courts among legal actors (and the subsequent imposition of legal criteria to evaluate their work) does not rule out the possibility that they can (and often do) play a political function. As already stated, all legal decisions have certain outcomes, but Constitutional Courts, due to the task assigned to them in the constitutional architecture, make their decisions by looking to the legal outputs, namely the constitutionality or not

[65] From a legal theoretical perspective, this different kind of effect (output as legal and outcome as non-legal) can be considered a consequence of the more general distinction between the normative and social functions of the law. Joseph Raz, 'On the Functions of Law' in AWB Simpson (ed), *Oxford Essays in Jurisprudence (Second Series)* (Clarendon Press 1973) 280.

[66] E.g., Lopeman (n 7) 3–5 (as to the idea that judicial activism is primarily "intentional activism"). cf Christopher H Schroeder, 'Causes of the Recent Turn in Constitutional Interpretation' (2001) 51 Duke Law Journal 307, 352–353 (as to the difficulty to disconnect the legal reasoning leading to the legal outputs from the desired non-legal outcomes when it comes to constitutional interpretation).

of certain provisions. As stated by a legal scholar, "[c]ourts legislate, but that does not make them legislatures".[67]

Justices sitting on the highest benches certainly can (and often do) have a political agenda, but they are still forced to confront it with the legal system and its dominating principles. An inverted example can be seen among the members of parliament. They are unquestionably political actors with a clear political agenda, but they still sometimes play a very relevant legal function, and this is done in accordance with a specific legal agenda, that is, in accordance with legal criteria as to how the legal system or part of it should look. For example, this functional leaning into the legal world may happen when members of parliament in a special committee evaluate the legal limits of criminal liability attached to the highest position of the state, such as the President of the Republic or the Prime Minister.

Moreover, and connected to the latter, the legal features of Constitutional Courts are traceable to the fundamental ideology shaping their work. Justices working in Constitutional Courts operate in an environment which, though with many political passersby, has a primary legal task: to be the guardian ensuring that the law-making taking place in a certain state is done (or that a conflict among the highest public authorities is settled) in accordance with the highest rules fixed in the constitutional documents or fundamental laws. This task of Constitutional Courts is legal, in the sense that it consists of dealing with legal rules. When justices sit on the bench, they are assigned the primary task of checking the 'constitutionality' of certain legal rules: they ought to evaluate whether, from a legal discourse perspective (e.g., with the traditional rules regulating the legal reasoning), such legal rules can fit (or not) into the legal system as designed in the constitutional documents or fundamental laws. Obviously, justices are often well aware of the indirect political effects of their decisions (outcomes), an awareness that sometimes affects their settling on a certain solution instead of another. Regardless of any hidden agenda behind a certain decision, however, justices are always forced to "squeeze" their politically motivated decisions into boxes of legal justification, to

[67] Herbert M Kritzer, 'Martin Shapiro: Anticipating the New Institutionalism' in Nancy Maveety (ed), *The Pioneers of Judicial Behavior* (University of Michigan Press 2003) 409. See also Neil MacCormick, *Questioning Sovereignty* (OUP 1999) 11–15 (as to the fundamental ontological difference between the legal and political discourses); Luhmann (n 11) 162–165. cf Stone Sweet (n 7) 61; Jerold Waltman, *Principled Judicial Restraint: A Case against Activism* (Palgrave Pivot 2015) 58–61 (with a critical perspective as to the negative effects of introducing the legal paradigms employed by the judiciary into the political discourse).

whose fundamental principles and ways of reasoning the justices ought then to sacrifice (in case of conflicts) their political programs.[68]

In other words, to keep their legitimacy in the community, Constitutional Courts are always forced to speak the language of the law, not the one of politics, even in the cases when they aim to send political messages. As pointed out by Michel Foucault, language in modern society is power, and this feature persists even in the most "politicised" legal terminology: simply by classifying a political problem and a political solution into legal language, the justices are (consciously or unconsciously) choosing to impose on the issue the domain and limits set by the legal discourse.[69] Therefore, the work of the Constitutional Courts should be evaluated accordingly, that is, by using legal criteria and, at the same time, by excluding from the evaluation process all features and limits set by other types of discourses – and, among them, the political one.

VII. Conclusion

In light of the debate on judicial activism at the constitutional level, that is, the involvement of Constitutional Courts in political law-making, this article has investigated from a legal theoretical perspective the issue of whether such courts should be considered primarily legal or political actors. While Part II underlined the importance of Constitutional Courts' 'activism' inside the general issue of judicial activism, Part III pointed out the reasons why Constitutional Courts in established Western democracies, despite their hybrid nature, can be seen as being legal actors, but with strong ties to the political arena. The final Parts IV and V then offered some legal theoretical considerations on why one should normatively impose upon these courts the label of 'legal actors', though they play a political function.

Going back the initial metaphor, we can see that the Leaning Tower of Pisa has stood the test of time and, despite leaning precariously towards the ground, it is still considered (and functions as) a 'tower'. As is done for this tower, one ought

[68] Gonçalo de Almeida Ribeiro, 'Judicial Activism and Fidelity to Law' in Luis Pedro Pereira Coutinho, Massimo La Torre and Steven D Smith (eds), *Judicial Activism:* An Interdisciplinary Approach to the American and European Experiences (Springer 2015) 36–40; Dieter Grimm, *Constitutionalism: Past, Present, and Future* (OUP 2016) 208. *cf* Alec Stone Sweet, 'The Politics of Constitutional Review in France and Europe' (2007) 5 International Journal of Constitutional Law 69, 72. E.g., Lawrence B Solum, 'The Supreme Court in Bondage: Constitutional Stare Decisis, Legal Formalism, and the Future of Unenumerated Rights' (2006) 9 University of Pennsylvania Journal of Constitutional Law 155, 160–176; Ronald Dworkin, *Justice in Robes* (Belknap Press 2006) 147–150.

[69] Michel Foucault, *The Archaeology of Knowledge and the Discourse on Language* (Pantheon books 1972), ch 2. See also Alan Hunt and Gary Wickham, *Foucault and Law: Towards a Sociology of Law as Governance* (Pluto Press 1994) 7–12 and 41–43.

constantly to monitor the stance of the Constitutional Courts in relation to the political terrain: their (natural) activism ought always to be under the scrutiny of exclusively legal criteria (and not, for instance, of political opportunity) and the entrance of the political world into the Constitutional Courts should always be limited to their function and not their structure (e.g., by further politicising the selection process of the justices). Otherwise, by leaning too much towards the political ground, Constitutional Courts, like the Tower of Pisa, run the risk of losing their structural integrity and, becoming simply rocks scattered across the political field, which would fundamentally modify the institutional landscape of a democracy.

Cambridge Law Review (2021) Vol VI, Issue i, 30–44

Positioning Indigenous Law in the Legally Pluralistic State of Canada

Frankie Young*

Abstract

The *Beaver v Hill* decision is a key legal decision in Canada that deals with the application of private international law to resolving a family law dispute involving Indigenous litigants. Chappel J, for the trial court, found that it was appropriate to apply private international law principles to resolve a private law matter where Indigenous litigants are concerned. On the contrary, Lauwers J, for the Court of Appeal, found that the section 35(1) claim raised by the respondent was the best means to resolve the matter. Without supplying a well reasoned analysis, Lauwers J found that the trial court erred in applying private international law principles on the grounds that Indigenous law and Aboriginal law are not considered foreign law. Lauwers J is correct that Aboriginal law is not foreign law. Indeed, these common law principles have evolved over time and are meant to regulate dealings between the state, third parties, in some cases, and Indigenous peoples. On the contrary, Indigenous law is wholly distinguished from Aboriginal law. Indigenous legal principles have existed since time immemorial and regulate the relationships within Indigenous communities. While all legal traditions are derived from the cultural norms within a community of citizens, for Indigenous communities these cultural norms have evolved through continuous interpretation by elders and law-keepers. These legal traditions are foreign to the common law just as the civil law is foreign to the common law. I argue that Lauwers J erred in finding that private international law principles should not be applied to resolve private law disputes that involve Indigenous litigants because he failed to recognize that, in keeping

* Assistant Professor at Western Law School, Western University, Canada. Thank you to the three anonymous reviewers, and Professors Ghislain Otis, Dwight Newman and John Borrows, whose helpful comments assisted in strengthening this article. Any errors remain my own. frankie.young@uwo.ca.

with choice of law principles, any legal order that is not considered the law of the forum is considered foreign law. Further, given the high bar to meet in asserting section 35(1) claims, it is disempowering to Indigenous Nations to assert that the section 35(1) claims are the only means for Nations to assert the application of Indigenous law.

Keywords: choice of law, indigenous law, legal pluralism, family law, private law

I. Introduction

On 12 October 2018 the Ontario Court of Appeal (ONCA) released its decision in *Beaver v Hill*,[1] involving a private family law dispute between two Haudenosaunee litigants and members of the Six Nations of the Grand River, Ms Beaver (the applicant) and Mr Hill (the respondent). This case raises larger and contentious issues around the applicability of Indigenous law to resolve private law disputes involving Indigenous litigants. One of the issues the ONCA was tasked with was determining whether the Superior Court had erred in applying conflict of law principles to find that Ontario courts, and not the Haudenosaunee Confederacy, had jurisdiction to hear the matter. The ONCA rejected the application of private international law principles and instead found that asserting a constitutional claim under section 35(1) is the most effective route to resolve jurisdictional disputes involving Indigenous litigants.[2] Notwithstanding, many section 35(1) claims have not had positive outcomes because of the high bar that Indigenous petitioners must meet to successfully make out a claim.

I assert that to resolve private law disputes, in some cases it will be appropriate to apply conflict of law principles to determine Indigenous jurisdiction or the applicability of Indigenous law. Presumptions that section 35(1) claims are the only recourse available to Indigenous litigants should be avoided. Furthermore, while the trial court engaged in a jurisdictional analysis because choice of law was not raised by Mr Hill, I contend that it is entirely appropriate for an Indigenous litigant to plead foreign law as a means to argue that, rather than the law of the forum, Indigenous law should be applied to resolve the dispute. Pleading foreign law as a means to assert law other than the law of the forum is well recognised in Canadian jurisprudence.[3] Because Indigenous laws in Canada have historically operated as separate legal orders long before European contact, an Indigenous litigant who

1 [2018] ONCA 816 (application for leave to appeal to SCC dismissed July 4, 2019).

2 The Constitution Act 1982.

3 For example: *Boulanger v Johnson & Johnson Corp* [2003] 64 OR (3d) 208 (Div Ct); *General Motors Acceptance Corp of Canada v Town and Country Chrysler Ltd* [2007] 88 OR (3d) 666; *Phillips v Ford Motor Co of Canada* [1971 2 OR 637 (CA); *Hunt v T&N plc* [1993] 4 SCR 289.

pleads Indigenous law as 'foreign' law is simply asserting that the decision-making process should be governed by an *alternative* set of laws rather than the laws of the forum in question.

II. Background

After a five-year relationship which produced one son, Ms Beaver, who lived off reserve with their son, and Mr Hill, who lived on reserve, experienced a breakdown in the domestic relationship. Ms Beaver made an application before the Ontario Superior Court (ONSC) for custody, spousal support and child support.[4] After initially responding to the Ontario court, Mr Hill subsequently gave notice that he was challenging the jurisdiction of the court and the applicability of Ontario law. He filed a Notice of Constitutional Question (which was amended numerous times and was still found to be deficient by both levels of court) to assert that, pursuant to section 35(1) of the Constitution, he had an Aboriginal right of self-government which was being infringed by the imposition of Ontario family law, and the infringement was not justified.[5] The key legal issue was whether jurisdiction should be decided via the application of the conflict of law principles respecting jurisdiction, or via the section 35 constitutional framework respecting the determination of Aboriginal rights claims.[6]

A. Position of the applicant

Ms Beaver sought a declaration that the ONSC had jurisdiction to hear the application, pursuant to section 97 of the Courts of Justice Act 1990, and under the common law rules respecting jurisdiction. She raised the traditional ground of attornment to contend that Mr Hill attorned to the jurisdiction of the court when he initially served and filed an Answer and Claim in reliance on Ontario legislation.[7] Ms Beaver's counsel also submitted that under the principles of private international law, *jurisdiction simpliciter* was established such that, even if the Haudenosaunee are considered to be sovereign peoples with their own laws, a real

4 *Beaver* (n 1) [1]-[2].

5 ibid [2]-[3].

6 ibid [48]. This framework was initially articulated in *R v Sparrow* [1990] 1 SCR 1075, expanded upon in *R v Van der Peet* [1996] 2 SCR 507 and affirmed in *Lax Kw'alaams Indian Band v Canada (Attorney General)* [2011] SCC 56. The claimant must first characterise the asserted Aboriginal right and then demonstrate the existence of the pre-contact practice, tradition or custom that was integral to the distinctive pre-contact Aboriginal society. The claimant must also demonstrate that the right is a continuation of a pre-contact practice. Next, a claimant must demonstrate that there has been an infringement of the right established. If an infringement is proven, the Crown has the burden to prove that it is justified.

7 *Beaver* (n 1) [27].

and substantial connection existed between the litigants and Ontario.[8] Ms Beaver asserted that although the court can decline to assume jurisdiction in cases where it is clear another forum is more appropriate to determine the outcome of legal proceedings (in accordance with the doctrine of *forum non conveniens*), no evidence was presented to support this position and the material facts were not pleaded appropriately.[9] Ms Beaver further claimed that individuals lack standing to assert a constitutionally protected Aboriginal right to self-government. Therefore, the general principles respecting jurisdiction should apply.[10]

B. Position of the respondent

Mr Hill asserted that jurisdiction ought to be determined via the application of the section 35(1) tests articulated by the Supreme Court of Canada (SCC) for the determination of Aboriginal rights and the justification of infringement framework.[11] Most notably, Mr Hill contended that the general common law rules on which Ms Beaver relied are intended to apply to "foreign litigants, legal processes and laws"; he maintained that these principles do not apply because he is not a foreigner and the laws of the Haudenosaunee are a part of Ontario law.[12] Finally, he alleged that section 35(1) constitutional protection should supersede (and not be rendered subordinate to) the law respecting jurisdiction.[13] Mr Hill argued that because jurisdiction goes to the heart of his argument — and he sought the jurisdiction of the Haudenosaunee Confederacy — the dispute could only be settled after a fair hearing of his Aboriginal rights case where evidence is presented on current and pre-European contact practices, customs and traditions of the Haudenosaunee.[14]

C. Judgement of the superior court of ontario

Chappel J rendered the decision for the Ontario Superior Court. To render a binding decision, it must be found that the court has jurisdiction over the parties to the litigation and the subject matter of the dispute.[15] Chappel J first considered whether the conflict of law principles were relevant in *intra-provincial* jurisdiction disputes between two Ontario citizens. She found that indeed they applied because

8 ibid [29].
9 ibid.
10 ibid [26].
11 ibid [2], [34].
12 ibid [34].
13 ibid [34].
14 ibid [37].
15 Stephen Pitel & Nicholas Rafferty, *Conflict of Laws* (2nd edn, Irwin Law 2016) 1.

Mr Hill asserted an "alleged right to be governed by a complete system of dispute resolution, adjudicative processes and laws for handling Family Law matters that is independent of Ontario's court system, processes and laws".[16] Because the claim raised the preliminary issue around which forum should hear the matter and which laws should apply to resolve the dispute, Chappel J found the conflict of law principles were intended to address such queries.[17]

While she recognised Mr Hill's asserted constitutionally protected Aboriginal rights added complexity to the analysis, she nonetheless noted that the SCC has emphasised that when constitutional values are at issue, conflict of law principles ought to be malleable and adapted to account for such values.[18] The ability to adapt the common law in the face of constitutional challenges — especially since inquiries around the relevant legal culture that should determine these *sui generis* rights are normative inquiries — permits an analysis that incorporates the conflict of law principles while protecting Aboriginal rights.[19] As such, she opined that the starting point should be the conflict of law principles respecting jurisdiction, while also factoring in the constitutional issues.

In analysing whether the ONSC had jurisdiction, Chappel J considered the related two-step test: (1) determination of whether the court has or can assume jurisdiction (jurisdiction *simpliciter*) and (2) if jurisdiction *simpliciter* is established, whether the court should decline to take jurisdiction.[20] Due to the section 35(1) constitutional challenge (and the attornment issue raised), Chappel J modified the approach. Rather than only considering the family law legislation governing custody and access issues, or the rules of court permitting the court to assume jurisdiction *simpliciter* at step one of the analysis, she considered whether the court should exercise its discretion to decline jurisdiction to allow the constitutional challenge to proceed in the ONSC.[21] This approach would presumably allow for a full assessment of the jurisdiction issue based on the section 35(1) framework regarding Aboriginal rights claims. Chappel J cautioned however that although Aboriginal rights are critical and must be protected, the court is not obliged to consider them in a vacuum. As such, if Aboriginal rights are not pled and advanced

16 *Beaver* (n 1) [53].

17 ibid.

18 ibid [54]; *Morguard Investments Ltd. v De Savoye* [1990] 3 SCR 1077; *Hunt* (n 3); *Tolofson v Jensen* [1994] 3 SCR 1022*; and* Van Breda v Village Resorts Ltd. [2012] SCC 17.

19 *Beaver* (n 1) [55]; *R v Van der Peet* [1996] 2 SCR 507; *R v Sappier* [2006] 2 SCR 686; *Delgamuukw v British Columbia* [1997] 3 SCR 1010; Mark Walters, 'British Imperial Constitutional Law and Aboriginal Rights: A Comment on Delgamuukw v British Columbia' (1992) *17* QLJ 350, 412-13.

20 *Beaver* (n 1) 57.

21 ibid [65]-[67].

in a timely manner, to promote reconciliation the court must balance the interests of all parties involved.[22]

Chappel J further considered the principle of constitutional restraint such that if a case can be decided on constitutional and non-constitutional grounds, it should be decided on non-constitutional grounds where possible. In fact, she found that a court is not compelled to rule on a constitutional argument simply because one is raised.[23] She went on to determine the factors that should be considered in deciding whether to decline jurisdiction. She reiterated that where one party relies on the doctrine of *forum non conveniens*, that party has the onus to prove that the proposed alternative forum (in this case the Haudenosaunee Confederacy) is more appropriate than the forum which the opposing party is asserting.[24] This doctrine ensures the litigants a process to efficiently resolve the issue of forum. The SCC has held that a party must show that, based upon clear connecting factors, an alternative forum is more appropriate and the court should decline to exercise jurisdiction on the basis of *forum non conveniens*.[25] Notably, Mr Hill did not raise the issue of *forum non conveniens*, nor did he assert a forum or process in which he sought to proceed. Nevertheless, Chappel J found that a prolonged and complex constitutional proceeding should not be incorporated into the doctrine of *forum non conveniens* to resolve a preliminary issue as to the applicable alternative forum, procedure and law.[26]

Chappel J found that while courts have a duty to protect Aboriginal rights there is no absolute obligation to allow a full hearing of such claims where, from the outset, there are fatal deficiencies in the pleadings.[27] She also determined that Mr Hill did not have standing to assert the claim to self-government as an individual. In finding that the amended constitutional claim failed to set out a reasonable claim or defence in law, she dismissed Mr Hill's amended answer without leave

[22] ibid [68].

[23] ibid [69]; Phillips v Nova Scotia (Commission of Inquiry Into Westray Mine Tragedy) [1995] 2 SCR 97; R v Lloyd [2014] BCCA 224; Peter Hogg, Constitutional Law of Canada (5th ed, Supplemented, Vol 2, Thomson Reuters 2016) chapter 59, page 22.

[24] *Beaver* (n 1) [70].

[25] *Van Breda* (n 19) [82].

[26] *ibid [70]-[71].*

[27] *Lax Kw'alaams* (n 6).

to amend and found that Ms Beaver's application for custody, spousal and child support would proceed under Ontario law.[28]

D. Judgement of the ontario court of appeal

Lauwers JA, writing for the ONCA, overturned Chappel J's finding that the conflict of law principles applied.[29] He held that it was an error of law for her to apply the conflict of law principles because "[a]boriginal rights or Indigenous law do not constitute 'foreign law', even conceptually".[30] There was no analysis provided for this finding. Rather, Lauwers JA focused on the relevant framework that deals with section 35(1) rights and the overarching constitutional principles that should be considered to assess Mr Hill's standing as an individual litigant asserting a constitutional claim to self-government.[31] Ostensibly, these principles underscore the *sui generis* nature of Aboriginal rights, the evolution of constitutional law in this regard and the importance of specific tests set out by the SCC when assessing section 35(1) claims. Despite the communal nature of Aboriginal rights, Lauwers JA found that the combined principles and the nature of the claim asserted demonstrate that Aboriginal rights are exercised by individuals — thus have both collective and individual aspects — and in appropriate circumstances individuals can assert Aboriginal or treaty rights.[32]

He found that it was premature to dismiss Mr Hill's constitutional claim because the interests at stake were critical and Mr Hill's section 35 claim could not be evaluated at such an early stage of the proceeding and on such a deficient record.[33] He held that it is incumbent upon the court to consider an amendment as a means to remedy an insufficient cause of action and so Mr Hill was given leave to appear before another Superior Court judge to amend his constitutional claim.[34] It

[28] *Beaver* (n 1) [128]-[129].
[29] ibid [17]-[18].
[30] ibid [17].
[31] ibid [28]-[34].
[32] ibid [36] citing Justice Lebel in *Behn v Moulton Contracting Ltd.* [2013] SCC 26 [33]; comprehensive analysis at [39]-[69].
[33] ibid [13]; *Spar Roofing and Metal Supplies Ltd. v Glynn* [2016] ONCA 296 [37].
[34] *Beaver* (n 1) [13]-[14].

is clear that Lauwers JA placed considerable weight on constitutional principles as a way for Indigenous litigants to have their legal rights recognised.[35]

III. Analysis

There is no doubt that the ability of Indigenous Nations to resolve legal disputes under traditional law has been the subject of much discourse in Canada. In July 2018, Canada released the *Principles Respecting the Government of Canada's Relationship with Indigenous Peoples*, which emphasise that "interactions between federal, provincial, territorial, and Indigenous jurisdictions and laws"[36] should be underpinned by the recognition of Indigenous Nations' inherent jurisdiction and legal orders. The Truth and Reconciliation Committee also advocates for the recognition and implementation of Indigenous legal systems as an act of reconciliation.[37] Further, the United Nations Declaration on the Rights of Indigenous Peoples affirms the right of Indigenous people to maintain their legal systems and customs.[38]

Canada is recognised for its cultural and legal diversity.[39] Arguably, Indigenous law — which *pre-dates* European contact by thousands of years — ought to be treated as a separate legal system from which laws are "freely chosen by persons desiring to do so".[40] Other countries, for example South Africa, recognise that weight ought to be given to customary law in resolving private law disputes involving Indigenous peoples, and the conflict of law principles are applied to determine the appropriate applicable law.[41] Lauwers JA's default position that favours the application of Canadian law to determine an Aboriginal right to

[35] ibid [78]. It is worth noting that the court took issue with the long drawn out proceedings of which both parties contributed to the procedural morass.

[36] Minister of Justice and Attorney General of Canada, 'Principles Respecting the Government of Canada's Relationship with Indigenous Peoples' (2018) <https://www.justice.gc.ca/eng/csj-sjc/principles-principes.html> Principle 4 (accessed January 25, 2021).

[37] Truth and Reconciliation Commission of Canada, Truth and Reconciliation (*Commission of Canada: Calls to Action*, 2015) < http://trc.ca/assets/pdf/Calls_to_Action_English2.pdf>, No 42, 45(iv) (accessed January 25, 2021).

[38] United Nations Declaration on the Rights of Indigenous Peoples, UNGA Resolution 61/295 (13 September 2007) UN Doc A/RES/61/295, articles 5, 24.

[39] Ghislain Otis, 'Individual Choice of Law for Indigenous People in Canada: Reconciling Legal Pluralism with Human Rights?' (2018) 8 UC Irvine Law Review 207, 213.

[40] Hadley Friedland, 'Navigating Through Narratives of Despair: Making Space for the Cree Reasonable Person in the Canadian Justice System' (2016) 67 UNBLJ 269, 13-14.

[41] Customary law is recognised in the Constitution of South Africa*:* C Rautenbach, *Introduction to Legal Pluralism in South Africa* (5th edn, LexisNexis 2018) 19; South African Law Commission, 'The Harmonisation of the Common Law and the Indigenous Law: Report on the Conflicts of Law' (*South African Law Commission Project 90,* 1999) 14-20; *Gumede v President of the Republic of South Africa* [2009] (3) SA 152 (CC) [22].

self-government presupposes that the application of this law is the best route for recognising distinct Indigenous legal systems. However, the rule of law asserted by Lauwers JA has historically been the same law that has upheld the dominant legal order or was applied to find Indigenous legal orders invalid.[42] Arguably, Lauwers JA's scepticism of the application of the conflict of law principles in this case was misguided and these principles should not be precluded as a propitious option for resolving private law disputes involving Indigenous peoples.

A. Pleading foreign law as an option

For litigants like Mr Hill, pleading foreign law, that is, the law from which a case should be resolved under, is a voluntary option.[43] Foreign law is any law other than the law of the forum — for example in an interprovincial context — and a litigant must plead proof of foreign law.[44] Typically, one of the parties must raise the choice of law for the court to even consider it, otherwise the legal matter is resolved under the law of the forum. Choice of law rules are those procedural rules that are applied by a court to determine which forum's law should apply to the matter at hand. Notwithstanding that the ONSC engaged in a jurisdictional analysis,[45] I argue that the choice of law rules could better serve the purposes sought here.

Albeit, Mr Hill did not plead foreign law and so the forum is not obligated to apply the choice of law principles, especially since foreign law is considered as a matter of fact and would have to be proven by Mr Hill.[46] There are likely several reasons why Mr Hill's counsel did not consider choice of law to resolve the issue at hand. First, the parties clearly misunderstood the nature of 'foreign law'. Mr Hill alleged that he is not a foreigner, and Haudenosaunee law is, rather than foreign law, *a part of the common law system* in Ontario.[47] Lauwers JA also asserted that both Aboriginal rights *and* Indigenous law should not fall within the scope of conflict of law principles because they "do not constitute 'foreign law', even conceptually".[48] Certainly Aboriginal rights are a part of Canadian constitutional law and are *not*

[42] John Borrows, *Questioning Canada's Title to Land: The Rule of Law, Aboriginal Peoples, and Colonialism' in Recovering Canada: The Resurgence of Indigenous Law* (University of Toronto Press 2002) 113.

[43] *Pettkus v Becker* [1980] 2 SCR 834 [854].

[44] *Boulanger* (n 3).

[45] It is trite law that the Superior Court has jurisdiction over constitutional and family law issues, and on appeal Mr Hill conceded the jurisdiction of the court. Beaver (n 1) [11]; *Canada (AG) v Law Society of British Columbia* [1982] 2 SCR 307 [326]-[327].

[46] *Hunt* (n 3) [308].

[47] *Beaver* (n 1) [34].

[48] *Beaver* (n 1) [17].

foreign law by any measure. Mr Hill's right to assert a constitutional claim is not disputed.

Wholly distinguished from Aboriginal rights, however, Indigenous laws have historically operated as separate legal orders long before European contact.[49] On its face, characterising Indigenous law as foreign law seems out of place. Indeed, Professor Karen Drake considers it ironic to characterise Indigenous traditions as foreign law within Indigenous traditional territories.[50] However, pleading 'foreign' law simply means asserting that the decision-making process should be governed by an *alternative* set of laws other than the laws of the forum in question.[51] In challenging the jurisdiction of the Ontario court, Mr Hill described the Aboriginal right he was relying on as "the Haudenosaunee right to be subject, sole [stet] and exclusively, to the family law and child support and parenting processes under Haudenosaunee law".[52] He further added that this right is "characterized as the exercise of an inherent right to self-government, which is recognized and affirmed as an Aboriginal and Treaty right by section 35 of the *Constitution Act, 1982*".[53] However, he went on to indicate that the Haudenosaunee have "not accepted the *imposition of Canadian laws* that touch on matters central to their society, namely governance and the application of provincial and federal statutory regimes that infringe on their core identity as a people".[54] Ostensibly, Mr Hill did not see the contradiction in asserting that Canadian law (the Aboriginal rights framework) should be applied to determine something as central to the Haudenosaunee as the right to self-government (thus the application of Haudenosaunee law), but denies that Canadian law applies to resolve the current family law dispute.

Moreover, when Mr Hill argued that applying conflict of law principles is offensive because he is not a foreigner and Indigenous law is not foreign law, rather a part of the common law of Ontario,[55] he directly contradicts his assertion that, rather than Ontario law, the separate and valid laws of the Haudenosaunee should apply. In fact, Mr Hill's Amended Answer asserted the existence of "a robust law, a dispute resolution system, which, among other things, determined how disputes within and between families were to be resolved" that "has been

[49] *Friedland* (n 42) [13]-[14].

[50] Karen Drake, 'Indigenous Oral Traditions in Court: Hearsay or Foreign Law?' in Karen Drake & Brenda L Gunn (eds) *Renewing Relationships: Indigenous Peoples and Canada* (Native Law Centre 2019) 3.

[51] Dicey and Morris, *Conflict of Laws* (15th edn, Sweet and Maxwell 2018); Pitel & Rafferty (n 16) 3; Adrian Briggs, *The Conflict of Laws* (4th edn, Oxford University Press 2019); *Beaver* (n 1) [52].

[52] *Beaver* (n 1) [6] ([29] of factum).

[53] ibid.

[54] ibid ([38] of factum).

[55] ibid [34].

practiced continuously since the time of contact with European settlers, despite the operation of other, colonial legal systems".[56] Although the claim was structured as a section 35(1) claim — one that had considerable defects[57] — what Mr Hill sought as an end result was for the court to apply Haudenosaunee law as the relevant law. Arguably, Haudenosaunee law is foreign to the Ontario rules governing these matters. This certainly appears to be a conflict of law issue because central to the function of the law of conflicts is that a pathway is provided to resolve controversies where law is being asserted other than the law of the forum.[58] Both Mr Hill and the ONCA struggled with the ability to reconcile Indigenous peoples and their law as having foreign elements to the forum.

Lauwers JA found that presumably Mr Hill would have no other means to assert Haudenosaunee laws and protocols, other than through a section 35 claim.[59] This is simply not true. Chappel J sought to reconcile the ways that private international law could intersect with constitutional principles to resolve these kinds of issues. Furthermore, Lauwers JA conceded that, under the current analysis, it is not clear whether Haudenosaunee law would entirely displace or simply modify Ontario family law such that Mr Hill's key assertions in his pleadings could, rather than supplant Ontario law, merely inform the process.[60] From this standpoint, there is no guarantee that simply proving a right to self-government and that the infringement by Ontario law is not justified would confer the right to have Haudenosaunee law automatically applied. Mr Hill would still have to deal with the fact that Ms Beaver and the child live off reserve, while Mr Hill lives on reserve. The relevant applicable law would still be an outstanding issue.

B. Indigenous legal traditions

Aside from the ambiguity around the nature of foreign law, the issue of how to treat Indigenous law as an effectual component of the multi-juridical Canadian legal system continues to be debated. Common law courts have erroneously

[56] ibid [3].
[57] ibid [89]; *Beaver* (n 1) [13].
[58] Gregory S Alexander, 'The Application and Avoidance of Foreign Law in the Law of Conflicts: Variations on a Theme of Alexander Nekam' (1976) 70(4) Northwestern University Law Review 602, 602.
[59] *Beaver* (n 1) [65].
[60] ibid [67].

treated Indigenous law as *evidence and fact* rather than as law.[61] This has not set good precedent for how to reconcile Indigenous law in the greater Canadian legal system, thus grave concerns have been raised around justice and fairness in cases involving Indigenous peoples. For instance, in *Coastal GasLink Pipeline Ltd. v Huson*, Wet'suwet'en customary law was not recognised as an effectual part of Canadian law, but could be considered as evidence in deciding a case; as such, the Wet'suwet'en peoples were found to be subject to the laws of British Columbia in resolving the legal issue in question.[62] However, numerous other legal decisions have recognised Indigenous legal orders as a part of the Canadian pluralistic legal system. In *Pastion v Dene Tha' First Nation*, the federal court found that "[i]ndigenous legal traditions are among Canada's legal traditions" and "form part of the law of the land".[63] Further, in *R v Marshall* the SCC cited, with approval, John Borrows: "[a]boriginal law should not just be received as evidence that Aboriginal peoples did something in the past on a piece of land. It is more than evidence: it is actually law".[64]

Barring a clear recognition of Indigenous legal orders as binding in their own right, similar to how Quebec civil law is recognised as having the same force and effect as the common law, Indigenous peoples (and ostensibly the courts) are likely to presume that the only recourse is to use the common law to assert section 35(1) claims. In fact, this case is an astounding example of how Mr Hill was seemingly

[61] *R v Van der Peet* [1996] 2 SCR 507 [84]-[91]; Drake (n 50) 17-21; Minnawaanagogiizhigook (Dawnis Kennedy), 'Reconciliation *W*ithout Respect? Section 35 and Indigenous Legal Orders' in Law Commission of Canada (ed) *Indigenous Legal Traditions* (UBC Press 2007), 87-89; Val Napoleon & Hadley Friedland, 'An Inside Job: Engaging with Indigenous Legal Traditions through Stories' (2016) 61*(4)* McGill LJ 725, 735.

[62] Coastal Gaslink Pipeline Ltd. v Huson Wet'suwet'en [2019] BCSC 2264 [128]. We see a similar line of reasoning in Logan v Styres, (1959) 20 DLR (2d) 416 where the ONSC found that the Haudenosaunee members of the Six Nations Indian Band were both under the protection of the laws of the land of Ontario, and were also subject to such laws. These decisions negate the legal rights of the Indigenous Nations at issue to be subject to their own laws.

[63] Pastion v Dene Tha' First Nation [2018] 4 FCR 467 [8]; see also Alderville First Nation v Canada [2014] FC 747 [26]; Connolly v Woolrich [1867] 17 RJRQ 75 (Qc Sup Ct); Re Adoption of Katie E7-1807 32 DLR (2d) 686 [36], [38]; Henry v Roseau River Anishinabe First Nation Custom Council [2017] FC 1038 [8]. For other cases that affirm the legitimacy of Indigenous law see Alexander v Roseau River Anishinabe First Nation [2019] FC 124; Campbell v British Columbia (Attorney General) [2000] BCJ No 1524. In fact, in the historical decision of *Connolly v Woolrich* [1867] 17 RJRQ 75 (Qc Sup Ct), the Quebec Superior Court held that a marriage entered into under Cree law could still be recognized under Quebec law. Further, in *Casimel v Insurance Corp. of British Columbia* [1993] BCJ No 1834 (QL) (BCCA), the BCCA considered the significance of customary adoption for the Carrier people (also known as Dakelh or Yinka Dene). The effective adoption of the late Chief Ernest Casimel by his grandparents was found to be legally binding. Both of these decisions recognized family law matters under customary law.

[64] *R v Marshall* [2005] SCC 43 [130].

of the view that his only recourse in law was to use the very system which he claims does not have jurisdiction over significant Indigenous legal matters to *make a substantive finding in law* that a section 35(1) constitutional right exists to be governed under Haudenosaunee customary law. Notwithstanding, the Haudenosaunee have historically disputed the legitimacy of the Canadian judiciary, asserting the right to be governed by their own laws.[65] Scholars and lawyers also express concern about whether claims under section 35(1) are a justiciable means to give effect to Indigenous people's laws.[66] Not only is there a high bar to meet but section 35(1) claims can take years to resolve.[67]

C. Choice of law rules

It can certainly be argued that choice of law rules are Canadian state law in the same way that section 35(1) is. Therefore, it is by virtue of a Canadian legal rule that foreign law could be received and given effect by Canadian law. However, the choice of law rules are largely procedural rules that determine which substantive law should apply. This is distinguished from applying substantive law to determine a critical legal finding related to Haudenosaunee self-government. This is also not to say that the choice of law principles would not have their own challenges for Indigenous parties asserting them. Central to the choice of law process is that a litigant is required to characterise the claim — which is critical to identifying the choice of law rule that determines the applicable law — and must precisely plead evidence of the content of the asserted law.[68] The connecting factors that favour the application of the pleaded law would need to be identified because choice of law principles are predicated on the principle of proximity such that the dispute is resolved according to the law most proximate to the dispute.[69] Mr Hill would thus need to deal with the fact that Ms Beaver and the child live off reserve, thus are connected to the law of Ontario, and spousal and child support are usually decided by the law of the forum. In *Kearney v Willis*, the court considered the choice of law under similar circumstances and, rather than applying the law where the

[65] *Sero v Gault* [1921] 64 DLR 327 (ONSC); Constance Backhouse, *Colour-Coded: A Legal History of Racism in Canada 1900-1950* (University of Toronto Press 1999) 117; *Logan v Styres* (1959) 20 DLR (2d)416.

[66] John Borrows, *Freedom and Indigenous Constitutionalism* (University of Toronto Press 2016) 27-40; Val Napoleon & Hadley Friedland, 'Indigenous Legal Traditions: Roots to Renaissance' in Marcus Dubber (ed) *Oxford Handbook of Criminal Law* (Oxford University Press 2014); Borrows (n 44).

[67] *Delgamuukw* (n 20). The trial lasted 374 days and new trial ordered due to a technicality.

[68] *Pitel & Rafferty* (n 16) 252.

[69] ibid 209.

respondent lived, the court applied the law of the forum because the child was born and raised there.[70]

Further, Mr Hill would have to overcome the fact that Ms Beaver disputed the existence of a specific "robust law" to deal with Haudenosaunee family disputes.[71] While the burden is on Mr Hill to provide evidence of the asserted law, if it is not clear to the court which rules from the Haudenosaunee legal system are to be applied to resolve the dispute, Mr Hill may not be successful.[72] Moreover, even if the law is proven by Mr Hill, Ms Beaver would want to opt out of the application of Haudenosaunee law. Certainly Indigenous litigants asserting Indigenous law and protocols may not prevail in every claim. However, the courts and Indigenous litigants should avoid presumptions that section 35(1) claims are the only recourse available to Indigenous litigants to assert Indigenous jurisdiction or the application of Indigenous law to resolve private law disputes. This has the effect of negating the choice of law process as a viable means to address issues of justice about *which legal system* should apply,[73] when the choice of law process clearly goes to the heart of a claim such as Mr Hill's.

This case raises unresolved questions that courts and governments will have to turn their attention to moving forward. Will courts in Canada give effect to legal decisions rendered under Indigenous law? How will the complexities of the numerous Indigenous, provincial and federal jurisdictions be reconciled? Given issues around access to justice in general, how will Indigenous litigants deal with pleading expert evidence under conflict of law principles? These questions are beyond the scope of this commentary. However, because recognising Indigenous law is a work in progress, a combination of mechanisms will be required to give effect to Indigenous law.

It is incumbent upon courts and governments to develop new understandings of what constitutes law to consider the unique position of Indigenous law in the development of the legal system in Canada. It should be noted here that Bill C-92, *An Act respecting First Nations, Inuit and Métis children, youth and families* is also in the process of being finalized in Canada. This legislation affirms the legitimacy of Indigenous law in that Nations will be empowered with developing policies and laws that flow from their particular histories, cultures, and circumstances. Eventually Indigenous law, recognized through this legislation's framework, will be applied to resolve these kinds of family law matters. While this is very good

[70] *Kearney v Willis* [2001] 15 RFL (5th) 96 (Nfld UFC).

[71] *Beaver* (n 1) [66].

[72] *Pitel & Rafferty* (n 16) 222-23.

[73] Ugljesa Grusic, *'Historical Development and Current Theories' in Paul Torremans et al (eds)* Cheshire, *North and Fawcett Private International Law* (15th ed*n*, Oxford University Press 2017) 37.

news for this area of law, it does not mean that conflict of law issues may not still arise. Conflict of law principles should not be precluded as a mechanism for giving effect to Indigenous legal orders to resolve private law disputes involving Indigenous peoples.

IV. Conclusion

This judgment has serious socio-political repercussions. Lauwers JA's reasoning failed to give effect to private international law principles as a means to resolve Indigenous litigants' private law disputes under Indigenous law. It was found that the pleadings played a critical role in defining the issues in this case.[74] In fact, the ONCA found that "the ramshackle way in which the constitutional claim was asserted and [was] being developed" did not give "justice to the seriousness of the claim".[75] This speaks to the difficulties in framing constitutional claims in this regard. Clearly, with the right guidance, a more effective constitutional claim could be pleaded. However, the choice of law principles are designed to resolve the kind of queries raised by Mr Hill.

Rather than reverting to Aboriginal rights claims by default, private international law could be a means for Indigenous peoples to assert Indigenous jurisdiction and choice of law in resolving private law disputes. Choice of law principles have been applied in legal disputes in Canada to recognise that, where appropriate, laws other than the law of the forum should be applied to resolve a particular legal dispute. Given the longstanding recognition of Indigenous laws as an effectual part of Canada's pluralistic legal traditions, it should not be out of the realm of possibility that an Indigenous claimant would want to assert the application of Indigenous laws to resolve a legal dispute. Perhaps in some cases, as is likely the case here, the Indigenous litigant will not have the strongest set of facts. However, the *Beaver v Hill* decision raises the larger issue of how Canada's legal system is going to deal with the interplay between *all* of Canada's legal traditions going forward. In the furtherance of justice and equity, measures ought to be taken to give effect to the rich Indigenous legal traditions of Canada's Indigenous peoples. Indeed, as was so eloquently noted by English scholar Cheshire:

> "[w]hen the circumstances indicate that the internal law of a foreign country will provide a solution more just, more convenient, and more in accord with the expectations of the parties than the internal law of England, the English judge does not hesitate to apply the foreign rules".[76]

[74] *Beaver* (n 1) [30].
[75] ibid [13].
[76] *Torremans* (n 75).

Cambridge Law Review (2021) Vol VI, Issue i, 45–59

Has COVID-19 Unlocked Digital Justice? Answers from the World of International Arbitration

Domenico Piers De Martino* and Katharina Plavec**

Abstract

The article aims to provide an overview of 'digital arbitration' one year after the beginning of the COVID-19 pandemic, with a view to supporting its wider implementation. In particular, the article illustrates some of the additional benefits online hearings confer and presents the legal framework, giving examples of how arbitral institutions around the world have adapted to the constraints the pandemic has imposed, with reference to their arbitration rules. In order to investigate the legality of remote hearings, examples of relevant provisions in the law of the seat of the arbitration, and the impact of the New York Convention, are also considered. In addition, the article briefly explores how virtual proceedings take place in practice, highlighting some of the key factors to be taken into account when conducting a hearing online. We conclude that a single, uniform and exhaustive answer on the legality of virtual hearings is not possible. This is because the answer is conditional on the position adopted in legislation across multiple jurisdictions and it requires a case-by-case approach. Nevertheless, in general, remote hearings are permissible under the New York Convention regime, and are not prohibited by the national

* Domenico Piers De Martino holds a Master in Law and Finance from the University of Oxford and is a Trainee Lawyer in international arbitration in Paris. Domenico.DeMartino.MLF2019@said.oxford.edu.

** Dr Katharina Plavec holds a PhD in international arbitration from the University of Vienna and a Master in Law and Finance (Distinction) from the University of Oxford. Katharina.Plavec.MLF2019@said.oxford.edu.

arbitration laws which were analysed in the article. Therefore, their increased adoption is foreseeable in the near future.

Keywords: international arbitration, dispute resolution, remote hearings, digital justice, COVID-19

I. Introduction

The global spread of the COVID-19 epidemic has severely restricted people's mobility.[1] Not surprisingly, the international arbitration community has responded by strongly advocating for the adoption of long-distance communication technologies, including the shift to online platforms to carry out pending proceedings,[2] while also pointing to the shortcomings of a virtual process. In order to assess whether the move to online hearings will be permanent, one first needs to analyse the legality of this form of 'digital justice'.[3]

An agreement to hold an online arbitration is recognisable and enforceable under the New York Convention, as well as compatible with the UNCITRAL Model Law and encouraged by the rules of different arbitral institutions (as will be shown below). Nevertheless, online arbitration entails new challenges, such as preserving overall procedural fairness. For instance, 'due process' is a concern, especially with respect to the cross-examination of witnesses and in circumstances where the principle of orality cannot be disregarded.[4] Therefore, it is necessary to take into consideration on a case-by-case basis not only the arbitration rules applicable, but also the domestic laws of both the seat elected by the parties for the dispute, and of the country in which enforcement is sought. It is then possible to establish the extent to which the parties can opt for an online hearing and, in the event of disagreement between the parties on the point, in which circumstances

1 For historical information and updates on social distancing measures relating to COVID-19, refer to the World Health Organization's official website at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019, accessed 22 November 2020.

2 See, for instance, Maxi Scherer, 'Remote Hearings in International Arbitration: An Analytical Framework' (2020) 37(4) Journal of International Arbitration 407, 407 et seqq.

3 While the concept of "digital justice" may be used to indicate different forms of digitalisation of the law and its process—many of which may also involve some degree of automation—in this work, by "digital justice" the authors intend to refer to the specific process of de-materialising the courtroom and holding hearings virtually, hence delivering justice to people through online spaces.

4 Yvonne Mak, 'Do Virtual Hearings Without Parties' Agreement Contravene Due Process? The View from Singapore' (Kluwer Arbitration Blog, 20 June 2020) <http://arbitrationblog.kluwerarbitration.com/2020/06/20/do-virtual-hearings-without-parties-agreement-contravene-due-process-the-view-from-singapore/> accessed 22 November 2020.

the Arbitral Tribunal may mandate such an online hearing without jeopardising the enforceability of the award.

The goal of this article is to provide an overview of 'digital arbitration' one year after the start of the pandemic. Section II outlines the problems associated with online hearings. Section III presents the legal framework by providing examples of how arbitral institutions around the world have adapted to the pandemic, also referring to their arbitration rules. The article then assesses relevant provisions which may be contained in the law of the seat of the arbitration and considers the impact of the New York Convention. Finally, Section IV briefly explores how virtual proceedings take place in practice, highlighting some of the key factors to be considered when conducting a hearing online, and Section V gives a conclusion concerning the progress of their implementation.

Despite the fact that most practitioners currently still prefer in-person or semi-remote hearings,[5] it is safe to assume that the new technological solutions which are now adopted will continue to be deployed once social distancing measures have ceased. This is primarily because there are strong incentives among the international community to make international commercial dispute resolution more expedited and less costly.[6] Secondly, national laws and the treaty network, which sustain the international arbitration system, enable the parties to validly opt for the implementation of online hearings and enforce the outcomes. Any potential shortcoming arising from this type of practice will turn into an opportunity for those lawyers who are able to master this new phase of the proceedings by adjusting their toolkit to an online environment, thus limiting the disruption caused by the pandemic, and offering their clients a new set of skills in the future.

II. The Status Quo

Digital tools have been used in international arbitration long before the COVID-19 outbreak. In particular, the correspondence among the parties takes place via email, and video or phone conferences have been employed for preliminary meetings, typically to discuss administrative aspects of the process (e.g., for the case management conference). This has been facilitated by the spread

[5] Gary Born, Anneliese Day, and Hafez Virjee, 'Empirical Study of Experiences with Remote Hearings: A Survey of Users' Views', in Maxi Scherer, Niuscha Bassiri, and Mohamed S Abdel Wahab (eds), *International Arbitration and the COVID-19 Revolution* (Kluwer Law International 2020) 138.

[6] For some of the efficiency benefits envisioned by the international arbitration community, see Mirèze Philippe, 'Offline or Online? Virtual Hearings or ODR?' (Kluwer Arbitration Blog, 26 April 2020) <http://arbitrationblog.kluwerarbitration.com/2020/04/26/offline-or-online-virtual-hearings-or-odr/> accessed 22 November 2020.

of increasingly sophisticated and reliable forms of digital signatures,[7] which certify the origin of the files exchanged and the identity of the people involved; this allows all of the paperwork to safely circulate digitally.[8]

However, hearings on the merits have so far seldom taken place via video conferencing. Arguably, the stage of the procedure which is still linked to physical interactions more than any other is the oral examination of witnesses. Generally speaking, virtual interrogations might be ineffective with regard to the assessment of their credibility. More specifically, typical cross-examination tactics, such as using surprise effects or engaging in prolonged eye contact, might become impracticable. When lawyers and arbitrators are not in the same room, it also becomes harder to monitor the witness' behaviour as they might receive unlawful interference during the examination or access documents they were not supposed to access, thus leading to a reduction in the formality of the proceedings.[9]

This has recently been acknowledged by the Federal Court of Australia. The shortcomings brought about by online trials were discussed in an exceptionally clear manner in a case pending before this Court during the first lockdown of early 2020. After addressing the objections raised by the Respondent, the Court issued an Order listing the principles which should be taken into account when deciding on the feasibility of a virtual hearing. These principles were grouped into "technological limitations; physical separation of legal teams; expert witnesses; lay witnesses, and in particular cross-examination; document management and trial length and expense".[10] This confirms that most of the challenges are due to how difficult it is for the parties to interact and communicate as they normally would, and suggests that there is scope for a trade-off between the thoroughness of the

[7] On the formalities required for the validity of written documents across jurisdictions, see Reinmar Wolff, 'E-Arbitration Agreements and E-Awards: Arbitration Agreements Concluded in an Electronic Environment and Digital Arbitral Awards' in Maud Piers and Christian Aschauer (eds), *Arbitration in the Digital Age: The Brave New World of Arbitration* (Cambridge University Press 2018) 151–181; Felipe Volio Soley, 'Signing the Arbitral Award in Wet Ink: Resistance to Technological Change or A Reasonable Precaution?' (Kluwer Arbitration Blog, 6 November 2020) <http://arbitrationblog.kluwerarbitration.com/2020/11/06/signing-the-arbitral-award-in-wet-ink-resistance-to-technological-change-or-a-reasonable-precaution/> accessed 24 November 2020.

[8] As a leading example, starting in September 2019, the Stockholm Chamber of Commerce has been providing electronic case management and file sharing facilities through its proprietary digital platform for the arbitral institution, the parties and the Arbitral Tribunal, available at <https://sccinstitute.com/scc-platform/> accessed 13 November 2020.

[9] For an example of a witness's examination and the relative challenges, see Chahat Chawla, 'International Arbitration During COVID-19: A Case Counsel's Perspective' (Kluwer Arbitration Blog, 4 June 2020) <http://arbitrationblog.kluwerarbitration.com/2020/06/04/international-arbitration-during-covid-19-a-case-counsels-perspective/> accessed 22 November 2020.

[10] See *Capic v Ford Motor Company of Australia Limited* (Adjournment) [2020] FCA 486.

process and its duration and cost.[11] Before considering these challenges further, it is crucial to determine the legal basis for assessing the legality of virtual hearings.

III. Legal Basis for the Legality of Online Hearings

With regard to assessing the legality of online hearings, multiple sources come into play in international arbitration. First of all, one must consider the law, especially procedural guarantees, at the seat of the arbitration, as well as laws applicable in the country of enforcement. These might trump the rules selected by the parties and are crucial to the validity and enforceability of the award. This means that for a full exploration of the legality of online hearings, a comparative analysis of the mandatory provisions governing arbitration proceedings across jurisdictions is necessary. The UNCITRAL Model Law aids the interpreter in this enterprise, given the harmonising effect that the Model Law has had on the arbitration laws of those countries adopting its provisions.[12] Secondly, the New York Convention is also included in the relevant framework as it lists the conditions for recognising and enforcing foreign awards, and is therefore paramount for those parties seeking to effectively enforce an award issued at the end of a partially or entirely virtual process.

Furthermore, regardless of whether the parties have opted for an *ad hoc* or institutional arbitration, the rules of arbitral institutions can provide additional 'authoritative' support by expressly contemplating a series of provisions concerning virtual hearings.[13] The parties are free to include some of these rules in their arbitration agreements when arbitrating *ad hoc,* or they find them by default in their agreements if they decide to file with an arbitration institution, even in those cases where the parties never negotiated any rules with regard to online hearings clearly.

[11] In this case, the Supreme Court concluded that the virtual process was likely to increase the overall costs of the proceedings, see ibid. Notwithstanding that this consideration might be questionable, it is not conclusive for the present study on international arbitration where the parties would have to otherwise bear the costs of appearing before a foreign court or Arbitral Tribunal. This is arguably more expensive than holding part of the process online, especially when witnesses or expert witnesses are also involved.

[12] See Nigel Blackaby and others, *Redfern And Hunter on International Arbitration* (6th ed, Oxford University Press 2015) 58–60.

[13] The distinction between ad hoc and institutional arbitration is illustrated at ibid 43–47.

The support currently provided by arbitral institutions encourages the parties to use online hearings and adopt specific provisions in their arbitration agreements.

A. Arbitral Institutions

Most arbitral institutions have provided guidelines on the conduct of proceedings during the current pandemic. For example, the ICC's 'Guidance Note on Possible Measures Aimed at Mitigating the Effects of the COVID-19 Pandemic' is particularly detailed.[14] Similar guidance was released by the Chartered Institute of Arbitrators[15] and the Vienna International Arbitral Centre.[16] Likewise, in an official communication in April 2020, the Milan Chamber of Arbitration invited the Tribunals to make every possible effort to carry out pending hearings via video or audio conference.[17] The website of Delos contains a comprehensive overview of checklists issued by various arbitral institutions as well as links to many of the webinars that have been conducted since the beginning of the pandemic.[18]

These leading arbitration institutions were also part of a joint communication which was released to encourage Arbitral Tribunals and parties to identify measures necessary to address the challenges arising from the pandemic and preserve the efficiency of arbitration proceedings. On that occasion, explicit reference to the use of digital technologies for working remotely was made.[19] These initiatives are clear evidence that online hearings are seen by leading arbitral institutions as a

[14] ICC Guidance Note on Possible Measures Aimed at Mitigating the Effects of the COVID-19 Pandemic (9 April 2020), <https://iccwbo.org/content/uploads/sites/3/2020/04/guidance-note-possible-measures-mitigating-effects-covid-19-english.pdf> accessed 22 November 2020.

[15] Chartered Institute of Arbitrators, Guidance Note on Remote Dispute Resolution Proceedings (8 April 2020), <www.ciarb.org/media/8967/remote-hearings-guidance-note.pdf> accessed 22 November 2020.

[16] Vienna International Arbitration Centre, A Practical Checklist for Remote Hearings (June 2020), <www.viac.eu/images/documents/The_Vienna_Protocol_-_A_Practical_Checklist_for_Remote_Hearings_FINAL.pdf> accessed 22 November 2020.

[17] Milan Chamber of Arbitration (14 April 2020) <www.camera-arbitrale.it/it/news/arbitrato-sospensioni-dei-termini.php?id=930> accessed 22 November 2020.

[18] See https://delosdr.org/index.php/2020/05/12/resources-on-virtual-hearings/, accessed 22 November 2020. For further information on this topic, see Patricia Louise Shaughnessy, 'Initiating and Administering Arbitration Remotely' in Maxi Scherer, Niuscha Bassiri, and Mohamed S Abdel Wahab (eds), *International Arbitration and the COVID-19 Revolution* (Kluwer Law International 2020) 27–48.

[19] The joint statement may be viewed at https://iccwbo.org/content/uploads/sites/3/2020/04/covid19-joint-statement.pdf, accessed 22 November 2020.

viable option in the context of the disruption caused by COVID-19, and possibly for improving arbitration in the future.

Moreover, when parties choose a particular set of arbitration rules, they are generally free to tailor them to their specific needs, filling any existing gap by directly contracting with each other on those aspects of the procedure which are not already regulated by the rules; or they can deviate from the non-mandatory parts of the rules. This would include adapting the chosen rules to the setting up of online hearings, if deemed necessary. In any case, if the parties have not agreed *ex ante* on the matter or are unable to reach an agreement during the proceedings, the Arbitral Tribunal may well mandate that the hearing be held online, as long as this is aligned with the arbitration rules agreed upon by the parties and not prohibited by the applicable law of the seat.

Under the rules of some of the most prominent arbitration institutions, granting extensive discretionary powers to the Arbitral Tribunals is the norm. For example — and without having the expectation of providing a universal picture — Article 17 of the International Rules of the Korean Commercial Arbitration Board[20] and Article 19.1 of the 2016 Singapore International Arbitration Center Rules[21] envisage this principle. Additional evidence for this point is provided by Article 13.1 of the 2018 Hong Kong International Arbitration Centre Administered Arbitration Rules,[22] where the discretion of the tribunal encompasses making use of technology as it considers appropriate, and Article 28.1 of the Vienna International Arbitral Centre 2018 Arbitration Rules, entitled "conduct of the arbitration".[23] Similarly, reference to the arbitrators' discretion is made at Article 23 of the Stockholm Chamber of Commerce 2017 Arbitration Rules,[24] and at Article

[20] "The Arbitral Tribunal shall conduct the proceedings in accordance with the Rules and, where the Rules are silent, any rules which the parties or, failing them, the Arbitral Tribunal may settle on".

[21] "The Tribunal shall conduct the arbitration in such manner as it considers appropriate, after consulting with the parties, to ensure the fair, expeditious, economical and final resolution of the dispute".

[22] "Subject to these Rules, the arbitral tribunal shall adopt suitable procedures for the conduct of the arbitration in order to avoid unnecessary delay or expense, having regard to the complexity of the issues, the amount in dispute and the effective use of technology, and provided that such procedures ensure equal treatment of the parties and afford the parties a reasonable opportunity to present their case".

[23] "The arbitral tribunal shall conduct the arbitration in accordance with the Vienna Rules and the agreement of the parties in an efficient and cost-effective manner, but otherwise according to its own discretion. The arbitral tribunal shall treat the parties fairly. The parties shall be granted the right to be heard at every stage of the proceedings".

[24] "The Arbitral Tribunal may conduct the arbitration in such manner as it considers appropriate, subject to these Rules and any agreement between the parties. In all cases, the Arbitral Tribunal shall conduct the arbitration in an impartial, efficient and expeditious manner, giving each party an equal and reasonable opportunity to present its case".

32 of the American Arbitration Association's most recent Commercial Arbitration Rules,[25] the latter of which specifically provides for the power of the tribunal to mandate the use of different instruments of communication. The discretion with which Arbitral Tribunals are usually entrusted by the parties arguably ought to include the option of holding a hearing online, as this interpretation enables the arbitrators to select the most appropriate method for conducting the hearing, adjusting for the circumstances of the case with which they are confronted.

In support of this argument, the ICC has updated its Arbitration Rules, which entered into force on 1 January 2021. These rules explicitly give the arbitrator the discretion to mandate that a hearing be held by remote means of communication. In detail, the revised form of Article 26.1 states that "the arbitral tribunal may decide, after consulting the parties, and on the basis of the relevant facts and circumstances of the case, that any hearing will be conducted by physical attendance or remotely by videoconference, telephone or other appropriate means of communication".[26] Article 19.2 of the London Court of International Arbitration (LCIA) Rules, which were revised in response to the Pandemic and came into force in October 2020, reinforces this trend by explicitly detailing that "as to form, a hearing may take place in person, or virtually by conference call, videoconference or using other communications technology with participants in one or more geographical places (or in a combined form)".

Therefore, the discretionary powers conferred by default on Arbitral Tribunals in accordance with the rules of some of the main arbitral institutions may well be interpreted as enabling the arbitrators to mandate an online hearing. Moreover, even when the rules are silent on this matter, there is a new trend of including in each set of rules a provision expressly conceived for this purpose. This is the case especially regarding the sets of arbitration rules which used to be silent on the adoption of alternative means of virtual communication, such as the rules of the ICC and LCIA.

The situation is even clearer in the UNCITRAL Model Rules which take into account precisely the most problematic phase of the process, namely that of the testimony of the witnesses. The text of the Model Rules states beyond doubt that it may occur entirely by way of a virtual hearing. More specifically, Article 28.4 of the UNCITRAL Rules empowers the Arbitral Tribunal to hold examinations through "means of telecommunication that do not require their physical presence at the hearing (such as videoconference)". This rule is premised on Articles 17.3 and 27.2 of the UNCITRAL Rules. These provide that there is

[25] American Arbitration Association Commercial Arbitration Rules and Mediation Procedures, Rule 32(a), (b) and (c).

[26] ICC Rules of Arbitration, Article 26.

usually no obligation to hold parts of the arbitration process orally, as written and signed statements constitute a default rule, unless one of the parties requested an oral hearing with the arbitrators' approval, or the parties agreed on a specific mode in the first place.

As many arbitration rules move in the direction of facilitating online hearings, the assessment of the adequacy of such a process should be done on a case-by-case basis by the arbitrators, adjusting for the complexity of the dispute. To that end, the adoption of solutions which are tailored to the dispute is certainly not impeded by virtual hearings, but is in fact enhanced. Allowing flexibility on the matter is the approach most consistent with the founding principles of international arbitration.

B. Law at The Seat of the Arbitration

In addition to the applicable arbitration rules, which in most cases have proven to be in favour of online hearings, particular regard has to be had to the law at the seat of the arbitration, the *lex arbitri*, as this will ultimately determine the smoothness of the process. According to Article 24.1 of the UNCITRAL Model Law, the Tribunal shall decide whether to hold oral hearings or whether the proceedings shall be conducted on the basis of documents and other materials. Nevertheless, if one party requests a hearing, the Arbitral Tribunal should hold it at an appropriate stage of the proceedings. The possibility of online hearings is not explicitly mentioned in the Model Law; however, to some extent the fact that they are not prohibited supports the argument in favour of their legality.[27] Furthermore, according to Article 18 of the UNCITRAL Model Law, the parties shall be treated with equality and each party shall be given a full opportunity of presenting his case. This sets a broad standard in determining when a process contemplated by the parties, or implemented by the Arbitral Tribunal, is respectful of due process.

Similar rules to Articles 18 and 24.1 of the UNCITRAL Model Law are, for instance, contained in the Austrian Arbitration Law and the German Arbitration Law.[28] In Italy, the Articles 816-*bis* and *ter* of the Italian Code of Civil Procedure contain general provisions as to the extension of the powers of the Arbitral Tribunal throughout the process and the collection of evidence respectively. Arbitrators have broad powers to conduct the arbitral proceedings, provided that they respect the determination of the parties, guarantee due process and comply with public policy.

[27] See also Erica Stein, 'Challenges to Remote Arbitration Awards in Setting Aside and Enforcement Proceedings', in Maxi Scherer, Niuscha Bassiri, and Mohamed S Abdel Wahab (eds), *International Arbitration and the COVID-19 Revolution* (Kluwer Law International 2020) 173.

[28] See § 598 of the Austrian Code of Civil Procedure and § 1047 of the German Code of Civil Procedure.

Hence, in most cases, arbitrators are permitted to set the rules for specific aspects of the procedure as they deem most appropriate.[29] In these cases, notwithstanding the lack of an explicit provision on virtual hearings, it can be argued that the Tribunal has discretion to conduct hearings by video conferencing as long as no party objects.

Turning again to the Federal Court of Australia, it may be noted that no legal provision against virtual hearings was cited in *Capic v Ford Motor Company of Australia* to which this article referred previously.[30] This suggests that, if Australia, Italy, and Austria are assumed to be sufficiently representative jurisdictions, then finding a legal barrier to the implementation and enforceability of virtual arbitrations at 'seat level' might be unlikely, mainly due to the lack in most countries of an express provision which engages with the matter. It can therefore be argued with greater confidence that the fact that Australian judges have questioned the unsatisfactory nature of performing a cross-examination by video[31] does not lead to the conclusion that this practice violates due process, nor that it is against the free determination of the parties.

Apart from Australia, Austria is one of the few jurisdictions in which a higher court has ruled on the legality of online arbitration hearings.[32] In the case before the Supreme Court, the respondent did not agree to conduct the hearing via video-conferencing and subsequently challenged the Arbitral Tribunal over its decision to proceed with an online hearing. The Austrian Supreme Court found that remote hearings are permissible under Austrian law.

First, the Supreme Court held that conducting the hearing via a video conference did not violate the mandatory duty to treat the parties fairly as contained in the Austrian Arbitration Act, since the parties were granted sufficient time to prepare for the hearing. Most importantly, the Court held that using video technology in arbitral hearings does not violate Article 6 of the

[29] Michelangelo Cicogna, 'Arbitration in Italy' (Lexology, 9 January 2019) <www.lexology.com/library/detail.aspx?g=8c6a9ef8-00d1-4613-92e0-1e0ef728bed4#:~:text=Under%20Italian%20arbitration%20law%2C%20the,the%20inaction%20of%20the%20parties.> accessed 22 November 2020.

[30] See *Capic v Ford Motor Company* (n 10).

[31] For example, see *Hanson-Young v Leyonhjelm* (No 3) [2019] FCA 645 [2] and *Capic v Ford Motor Company* (n 10).

[32] Austrian Supreme Court 23 July 2020, 18 ONc 3/20s; for an English summary and comment see: Maxi Scherer and others, 'In a 'First' Worldwide, Austrian Supreme Court Confirms Arbitral Tribunal's Power to Hold Remote Hearings Over One Party's Objection and Rejects Due Process Concerns' (Kluwer Arbitration Blog, 24 October 2020) <http://arbitrationblog.kluwerarbitration.com/2020/10/24/in-a-first-worldwide-austrian-supreme-court-confirms-arbitral-tribunals-power-to-hold-remote-hearings-over-one-partys-objection-and-rejects-due-process-concerns/> accessed 22 November 2020.

European Convention on Human Rights (ECHR), even in the absence of both parties' agreement. It stressed that Article 6 ECHR requires a trade-off between safeguarding the parties' right to be heard with the right to effectively pursue their civil rights. As a virtual hearing can save time and costs, especially in the time of a pandemic, it is, in the opinion of the Austrian Supreme Court, an effective and legal method combining efficient law enforcement with the right to be heard.

The Court further held that potential abuses concerning witness examinations cannot undermine the legality of video conferencing, emphasising that abuses such as witness coaching are also possible in regular hearings. It even found that virtual hearings offer further possibilities to prevent such abusive practices. For instance, witness examination could be recorded and, if there is a danger that a witness received private messages on their screen, he or she could be ordered to look directly into the camera. Moreover, it also explicitly rejected an argument based on the fact that the hearing was scheduled for 15h CET, i.e., outside the "classic working hours" for a key witness located in Los Angeles. It specifically stressed the fact that the parties had agreed to an arbitration seated in Vienna and that the witness's participation at an early hour of the day was in any case less burdensome than the alternative: travelling from Los Angeles to Vienna to be heard in person.

Finally, courts could also be called upon in setting aside proceedings. The law at the seat is crucial at this stage because it determines under what circumstances an award may be challenged. For example, under the UNCITRAL Model Law, an award may be set aside if a party was otherwise unable to present its case (Article 34(2)(a)(i)), or if the arbitral procedure was not in accordance with the agreement of the parties or the law at the seat (Article 34(2)(a)(iv)).[33] Whether elements of remote hearings may fall under these grounds for challenge remains to be seen. As the grounds largely mirror those of the New York Convention, this will be further discussed in the subsection below.

C. The New York Convention

An award rendered after a virtual hearing should be enforceable pursuant to the New York Convention. The New York Convention, in spite of its distant origin in 1958, generally allows the parties to enforce an award that is the result of an online procedure. This interpretation derives from the words of the Convention

[33] See Stein (n 27) 169.

from which it is clear that only a domestic mandatory provision can prevent the parties or the Arbitral Tribunal from lawfully opting for a process held remotely.

Article V lists the conditions under which the enforcement of an award may be denied. According to Article V(1)(d), an award may not be enforceable if the party against whom it is invoked proves to the competent authority that "the arbitral procedure was not in accordance with the agreement of the parties, or, failing such agreement, was not in accordance with the law of the country where the arbitration took place". This means that an arbitral award would not be enforceable if, for instance, the arbitration which led to it was decided on the basis of an essential piece of evidence which had been collected in breach of the procedure contemplated by the parties in their agreement. Hypothetically, collecting a deposition via video-conference might lead to such a breach where the parties had not agreed to it explicitly and one of the aggrieved parties, after being negatively affected by the evidence, decided to challenge the enforceability of the award on the basis of the lack of consent.[34]

Similarly, the second part of Article V(1)(d) of the Convention states that in the absence of an agreement between the parties on the process, an award may not be enforceable when it has been issued in conflict with the procedural rules which are applicable in the seat of the arbitration. It may be inferred that the Arbitral Tribunal can mandate or restrict virtual activities — as the case may be — in those situations in which it is interpreting the national law of the country of the seat. In addition, utmost care must be given to how the judiciary in the place of enforcement interprets Article V of the New York Convention, as this will vary across contracting jurisdictions.

Therefore, in order to reduce the risk that an award is deemed to be unenforceable, it is recommended that lawyers agree expressly with their counterparty at the outset of or during the arbitral proceedings that hearings, and in particular cross-examinations, can be conducted virtually. This consent on the procedural rules trumps hypothetical national restrictions as long as they do not constitute mandatory law, and avoids putting the decision to the discretion of the Tribunal in case of disagreement. An illustration of such a clause is provided by the Milan Chamber of Arbitration, by which spontaneously or following the demand of one of the parties, the Arbitral Tribunal "may schedule a single hearing for the taking of evidence and a final discussion, […] held by videoconference, telephone

[34] The same point is made in Roberto Argeri and others, 'The Milan Chamber Of Arbitration Adopts New Measures In The Wake Of COVID-19 Pandemic' (Mondaq, 9 August 2020) <www.mondaq.com/italy/arbitration-dispute-resolution/971640/the-milan-chamber-of-arbitration-adopts-new-measures-in-the-wake-of-covid-19-pandemic-> accessed 22 November 2020.

or similar means of communication".[35] Equally, Article 19.2 of the LCIA Rules and 26.1 of the new 2021 ICC Rules constitute perfect examples.

Finally, parties might also argue that a virtual hearing violates their right to be heard or their right of equal treatment. These rights are both encompassed by Article V(1)(b) of the New York Convention. However, if both parties are given the same opportunity to present their case virtually and no technical issues occur, it is unlikely that this ground will be invoked successfully.[36]

In short, while many arbitral institutions have taken steps to encourage virtual hearings, the law at the seat of the tribunal and the New York Convention must also be taken into account. The present analysis suggests that neither the New York Convention nor laws based on the Model Law prohibit online hearings. Nevertheless, it is still advisable for parties to explicitly agree on this possibility beforehand. This agreement should consider the technical and practical implications of online settings, as these aspects may vary depending on the circumstances of the case and the location of the parties involved.

IV. Factors to be Considered when Conducting Online Hearings

While the international community reaches a consensus on the legality of online international arbitration, the potential problems arising from this new form of procedure call for lawyers to adapt their techniques and agree with the counterparty *ex ante* what rules shall apply to a cross-examination, for instance, in case a dispute subsequently arises. The factors which lawyers must take into account in advance include the time zones in which the parties and witnesses are based and their access to a stable and secure internet connection. Given that most national laws are currently silent on the legality of virtual proceedings and any additional requirements for the enforceability of the subsequent awards rendered by the Arbitral Tribunals, these elements should also inform the Tribunal's decision on whether or not they should impose an online hearing during the arbitration process. In fact, these factors potentially affect the overall procedural fairness, as well as substantive fairness as a consequence, thus determining when a virtual hearing is advisable for all parties involved.

Leading international arbitration institutions, such as the International Centre for Settlement of Investment Disputes (ICSID), are contributing by making available the IT tools necessary for a fully remote process.[37] Likewise, the

[35] See Article 5(5) of Annex D of the Arbitration Rules of the Milan Arbitration Chamber on simplified arbitration which are now applicable, since the 1st of July 2020.

[36] See the detailed analysis of Scherer (n 2) 439 et seqq.

[37] See the ICSID's case administration services on its official website available at https://icsid.worldbank.org/services/arbitration/case-administration, accessed 22 November 2020.

American Arbitration Association is offering a virtual hearing support service.[38] This is extremely important in terms of limiting any opportunistic arbitrages in the selection of the software deployed during the process, and absorbing some of the costs that otherwise would need to be borne by the parties. Accordingly, for online hearings to work, virtual spaces need to be cyber-secure and completely impenetrable with regard to potential hacks or privacy breaches aimed at obtaining confidential information.[39]

Different locations of hearing participants and choosing the right software are only two examples of practical issues that must be addressed at the outset of the arbitration. Thus, it is highly advisable that the Tribunal addresses these practical concerns in the form of a Procedural Order.[40] In spite of the technical support which a virtual hearing requires in order to be effective, to date they will cost between £3K and £5K per day in complex cases, depending on the number of participants, the composition of the legal team and the members of the Arbitral Tribunal.[41] Hence, this type of hearing is still likely to be cheaper than equivalent in-person hearings. This, in turn, constitutes a strong incentive for firms involved in international arbitration to continue developing the practice, and it is an additional element to consider when giving legal advice on the international dispute resolution options available to clients.

V. Conclusion

In conclusion, a single, uniform and exhaustive answer on the legality of virtual hearings is not possible. This is because the answer is conditional on the

[38] See https://go.adr.org/covid-19-virtual-hearings.html, accessed 13 November 2020.

[39] Cyber security and data protection challenges are not new to the world of arbitration: see Gerald Leong, 'How Do You Deal With Data Protection and Cybersecurity Issues In a Procedural Order?' (Kluwer Arbitration Blog, 19 February 2020) <http://arbitrationblog.kluwerarbitration.com/2020/02/19/how-do-you-deal-with-data-protection-and-cybersecurity-issues-in-a-procedural-order/> accessed 22 November 2020. However, the impact of the COVID-19 Pandemic, which is undoubtedly going to increase parties' reliance on remote forms of procedure, will exacerbate those problems and require more thought about the possible solutions. For example, see Gian Paolo Coppola and Marco Imperiale, 'Between Cybersecurity and Arbitration In Times Of Coronavirus Warnings, Suggestions And New Frontiers' (Mondaq, 5 August 2020) <www.mondaq.com/italy/security/972958/between-cybersecurity-and-arbitration-in-times-of-coronavirus-warnings-suggestions-and-new-frontiers> accessed 2 September 2020.

[40] For an example of such a Procedural Order, see Niuscha Bassiri, 'Conducting Remote Hearings: Issues of Planning, Preparation and Sample Procedural Orders', in Maxi Scherer, Niuscha Bassiri, and Mohamed S Abdel Wahab (eds), *International Arbitration and the COVID-19 Revolution* (Kluwer Law International 2020) 105–120.

[41] See Emily O'Neill and Mehdi Mellah, 'Hear us out: in-house litigators and the future of virtual hearings', (The Law Society, 28 October 2020) <https://www.lawsociety.org.uk/topics/in-house/the-future-of-virtual-hearings> accessed 22 November 2020.

position adopted in legislation across multiple jurisdictions. Parties are well advised to agree on the possibility of holding hearings virtually, either in their arbitration agreement or at a later stage. This choice seems more than advisable in a scenario where arbitrators from different nationalities are part of numerous tribunals spread across the globe, and where international lawyers are assisting several clients in different *fora* simultaneously. Moreover, this solution is permissible under the New York Convention regime, and also is not prohibited by the national arbitration laws taken into consideration. In line with the powers typically conferred on Arbitral Tribunals, in the absence of an agreement between the parties, the arbitrators should decide whether to mandate a virtual hearing, depending on the complexity of the case, its seat, and the specific situation of the parties. The more remote hearings are experienced during the COVID-19 pandemic, the more likely it is that 'digital' international arbitration will become a common option.

As a result, online hearings will often constitute a more efficient route for parties involved in international arbitration. This will benefit the clients of the most adaptable lawyers, who are able to perform effective cross-examination through a screen, or draft in their clients' agreements the relevant clauses. Hopefully, the transition to online arbitration will continue to be facilitated by international institutions, which may make it an available option in their sets of rules, starting with less complex disputes. This will help arbitration keep its competitive advantage over other alternative international dispute resolution methods in the future.

Cambridge Law Review (2021) Vol VI, Issue i, 60–96

Illegal and Inappropriate Evidence in International Investment Law: Balancing Admissibility

Aleksander Kalisz*

Abstract

The question of the admissibility of illegal or inappropriate evidence tests the limits of procedural flexibility of the arbitral process. Balancing admissibility requires a case-by-case approach. Tribunals will have to balance (or 'weigh') the substance of such documents with procedural fairness and general principles of law. In other words, the relevance of the evidence is weighed against the adverse and unfair effect that admission would have on the opponent. From an empirical perspective, reliance solely on the substance of the evidence rarely succeeds in outweighing procedural fairness. Exceptionally, however, publicly available documents, such as diplomatic cables leaked by WikiLeaks, have better chances of being admitted. The severity of the wrongfulness or unfairness may always tilt the balance in the opposite direction. Tribunals also unconditionally resist the admissibility of legally privileged documents. In any case, attempts to admit tainted evidence do not leave the opponent unprotected. The doctrine of equality of arms, good faith, and, debatably, the principle of clean hands safeguard them against unfairness. Finally, arbitrators have tools to tilt the scales of admissibility if the evidence is highly relevant. They may draw on the coercive powers of domestic courts through

* Commercial dispute resolution paralegal and future pupil barrister at CANDEY in London, akalisz@candey.com. I am grateful to the anonymous reviewers for their comments on earlier drafts. Any errors that remain are my own.

judicial assistance or order the production of documents to level the playing field for both parties.

Keywords: investment arbitration, international law, evidence, admissibility, procedural fairness

I. Introduction

One of the most eagerly cited advantages of arbitration is the flexibility of the process compared with litigation.[1] Tribunals generally have broad freedom to determine the procedural aspects of their cases. Despite clear advantages to the efficiency of proceedings, this flexibility can become a double-edged sword. Admissibility of evidence is one example. Arbitral tribunals, free from the requirements of civil procedure rules, might feel inclined to consider evidence that is inadmissible under domestic laws or *vice versa*. The treatment of such tainted evidence is further complicated by investment law being nested at the crossroads of public and private international law, and the principles from both influence the findings of tribunals.[2] The subject is particularly complex when the investor-sovereign State relationship is added to the discussion. Nonetheless, even in such complex circumstances, there must exist some principles on the admissibility of evidence to guide the tribunals.

This article analyses a narrow area of admissibility of evidence in investment arbitration — namely, the admissibility of illegally and inappropriately obtained evidence. It is clear that the process by which such tainted evidence is admitted is a weighing or balancing exercise — balancing the substantive relevance of the evidence with procedural fairness. The tainted evidence might be, after all, highly relevant to the dispute. On the other hand, the methods by which the evidence was procured may have been illegal or inappropriate. States have vast intelligence services, military technologies, and spying techniques to assist them. Investors, on the other hand, might be global corporations that are far more powerful and wealthy than some of the less economically developed respondent States. Such considerations of the balance of powers would fall into the procedural fairness analysis. In the end, tribunals balance these two considerations in deciding admissibility. This article takes a closer look at this process.

This article relies heavily on case law. The question asked is whether a common test for admissibility can be inferred from arbitral decisions, given that

1 William Park, 'Two Faces of Progress: Fairness and Flexibility in Arbitral Procedure' (2007) 23(3) Arbitration International 499, 499.

2 Andrea Brojklund and others, 'Investment Law at the Crossroads of Public and Private International Law' in August Reinisch, Mary Footer and Christina Binder (eds), International Law and… *Select Proceedings of the European Society of International Law* (Hart Publishing 2016) 151.

no clear test has been laid down in the applicable procedural rules or treaties. In addition, the article considers the procedural principles enshrined in Bilateral Investment Treaties (BITs), arbitration rules, and rules on the taking of evidence. This article focuses on the International Centre for Settlement of Investment Disputes (ICSID) Convention and Arbitration Rules and the United Nations Commission on International Trade Law (UNCITRAL) Arbitration Rules since they are the most widely used procedural rules in investment law. Case law is relevant because, although there is no doctrine of precedent in investment law, tribunals are prompted to follow a harmonious interpretation of international law and previous cases are clearly deemed highly authoritative.[3] In addition, the 2020 International Bar Association (IBA) Rules on the Taking of Evidence (IBA Rules) as well as the 2018 Rules on the Efficient Conduct of Proceedings in International Arbitration (Prague Rules) will be considered. They are frequently referred to by arbitral tribunals, despite being non-binding by themselves.[4]

The rationale for the research originates from the fact that rules on the admissibility of illegal and inappropriate evidence are scattered. Tribunals appear to lack a systematic approach to the issue and hence its resolution has been taken on a case-by-case basis. The situation is similar within the jurisprudence of the International Court of Justice (ICJ) and other international courts. This article hence considers whether any general tribunal practice may emerge from cases, hinting at the considerations which would or should be taken into account by future tribunals in admitting or rejecting tainted evidence. This is a complex question. Hence, the article takes a broad approach to the narrow issue of the admissibility of illegal and inappropriate evidence in investment arbitration.

Firstly, the article briefly discusses the ability of arbitral tribunals, which are not criminal courts, to analyse matters of illegality and impropriety associated with tainted evidence. Investment tribunals are arguably not created for that purpose, so this question of arbitrability deserves a mention.

Secondly, the article analyses the considerations for the balancing exercise. In particular, the relevant arbitration rules as well as case law are considered. Arbitration rules are relevant because they contain the framework of the tribunals' procedural powers, granted to the tribunals by the consent of States or party agreement. The extent of wrongfulness associated with admitting evidence

[3] AES Corp*oration v Argentina*, ICSID Case No ARB/02/17, Award (26 April 2005) [17]–[33]; Saipem *SpA v Bangladesh*, ICSID Case No ARB/05/07, Decision on Jurisdiction and Recommendation on Provisional Measures (21 March 2007) [67].

[4] See Cambodia Power v Cambodia, ICSID Case No ARB/09/18, Decision on the Claimant's Application to Exclude Mr Lobit's Witness Statement and Derivative Evidence (29 January 2012) [1]; Hrvatska Elektroprivreda *DD v Republic of Slovenia*, ICSID Case No ARB/05/24, Order Concerning the Participation of Counsel (6 May 2008) [19].

remains an important consideration for this element. Therefore, the different types of wrongfulness in international law are discussed, many of which are neither legal nor illegal when referring to the conduct of sovereign States. The more wrongful the conduct of one party, the less likely it is that their tainted evidence will be admitted.

Thirdly, the admissibility of tainted evidence likely stems from domestic laws. Hence, another consideration is the issue of admissibility of illegal and inappropriate evidence in domestic legal systems for comparative purposes. The discussed English and Austrian laws are most familiar to the author and illustrate the approach of common law and civil law traditions respectively. United States federal law, on the other hand, might reflect a general principle of law[5] and hence could indicate the direction of future developments.

In section IV, the article turns to procedural principles and how disadvantaged parties may be protected by investment law from tainted evidence being introduced against them. These considerations are used by tribunals if an imbalance is created by the new evidence. In such situations, investment law and general public international law might step in. That is because tribunals have to engage with broad international law considerations in ruling on admissibility, including particularly three principles of relevance: equality of arms, good faith, and clean hands.

The final section of the article analyses the limited tools tribunals can use to preserve the fairness of the arbitral process. Most relevant to the subject are judicial assistance requests and document production orders.

II. Arbitrability of Illegal and Inappropriate Conduct

The preliminary question is whether arbitral tribunals are a competent forum to address the impropriety or criminality of evidence. If illegality can be considered as part of the weighing exercise, this suggests that arbitrators have to engage with a task similar to domestic criminal courts. As will be seen, this is particularly true for circumstances of corruption.

The arguable function of investment law and investment tribunals is the protection of international trade. Mourre opines that arbitrators are "natural guardians of ethics and good morals in international commerce" and are "better placed than national judges to combat international fraud".[6] Although he refers to commercial arbitration, the statement is even truer for investment tribunals.

[5] Statute of the International Court of Justice, Article 38(1)(c).

[6] Alexis Mourre, 'Arbitration and Criminal Law: Jurisdiction, Arbitrability and Duties of the Arbitral Award' (2009) 19 International and Comparative Perspectives, International Arbitration Law Library 207, 207.

BITs are aimed primarily at promoting transnational trade, which should also be the ultimate goal of investment arbitration. As a result, although tribunals are not criminal courts, they may consider civil law consequences of criminal conduct.[7]

One of the most dominant types of unlawful conduct is corruption. It was addressed in detail in the *World Duty Free v Kenya* arbitration, where the tribunal stated that "bribery or influence peddling, as well as both active and passive corruption, are sanctioned by criminal law in most, if not all, countries".[8] In this case, the arbitrators did find evidence of corruption following a detailed analysis. The question of whether the prohibition of corruption constitutes a general principle of law is a separate discussion, but clearly it amounts to a violation of international public policy that tribunals enforce.[9] The tribunal quoted Judge Lagergren, who described international public policy as follows:

> "[w]hether one is taking the point of view of good government or that of commercial ethics it is impossible to close one's eyes to the probable destination of amounts of this magnitude, and to the destructive effect thereof on the business pattern with consequent impairment of industrial progress. Such corruption is an international evil; it is contrary to good morals and to an international public policy common to the community of nations".[10]

The *World Duty Free* case hence directly applied this analysis to investment arbitration, with a particular emphasis on corruption. Investment tribunals hence not only *may* consider illegal or inappropriate conduct, but should in fact do so. Bonifatemi adds that the issue of jurisdiction of tribunals in analysing corruption is a "non-issue".[11]

In the later case of *EDF v Romania*, the tribunal agreed with this conclusion, applying it to considerations of the admissibility of evidence.[12] In the case, corruption

[7] Dragor Hiber and Vladimir Pavic, 'Arbitration and Crime' (2008) 25(4) Journal of International Arbitration 461, 462.

[8] *World Duty Free v The Republic of Kenya*, ICSID Case No ARB/00/7, Award (4 October 2006) [142].

[9] ibid [138-41].

[10] J Gillis Wetter, 'Issues of Corruption before International Arbitral Tribunals: The Authentic Text and True Meaning of Judge Gunnar Lagergren's 1963 Award in ICC Case No. 1110' (1994) 10(3) Arbitration International 277, 294.

[11] Yas Bonifatemi, 'The Impact of Corruption on "Gateway Issues" of Arbitrability, Jurisdiction, Admissibility and Procedural Issues' in Domitille Baizeau and Richard Kreindler (eds) *Addressing Issues of Corruption in Commercial and Investment Arbitration* (ICC 2015) 18.

[12] EDF (S*ervices) Limited v Romania*, ICSID Case No ARB/05/13, Award (8 October 2009) [221].

was unsuccessfully argued by the claimant. Since the allegations concerned persons at the height of the Romanian Government, the tribunal pointed at the high standard of proof for such allegations.[13] In *Yukos v Russia*, the WikiLeaks evidence proved the misconduct of the respondent towards the claimant's auditors. These cases point towards the interplay between criminal considerations by the tribunals and the admissibility of evidence. The two frequently appear simultaneously and cannot be disentangled. Finally, in the *Awdi v Romania* arbitration, the tribunal used the criminal law language of a "presumption of innocence"[14] as a starting point for tribunals in assessing the culpability of parties for criminal allegations.

III. Weighing Exercise

Dolzer and Schreuer state that evidence in arbitration consists of documents, witness testimonies, and expert opinions.[15] The admissibility of such evidence is covered by the arbitration rules applicable to the dispute. Those would be mentioned explicitly in the BIT, subsequent party agreements or tribunal decisions, thus rendering them binding.

The ICSID Arbitration Rules stipulate in Article 34(1) that "the Tribunal shall be the judge of the admissibility of any evidence adduced and of its probative value". The Rules therefore leave wide discretion to the tribunal in considering factors for admissibility. Article 34(1) also mentions "probative value", prompting the arbitrators to look at the substance and usefulness of the evidence.

Article 25(6) of the UNCITRAL Arbitration Rules reads: "[t]he arbitral tribunal shall determine the admissibility, relevance, materiality and weight of the evidence offered". Caron and Caplan in the Commentary to the UNCITRAL Rules state that admissibility under the Article is "liberal pursuant to the spirit and practice", with the only exceptions being the evidence's "relevance, materiality and weight".[16] It should be noted that the passage does not explicitly mention that the manner in which the evidence was obtained is relevant, nor does it mention the legality or appropriateness of the evidence as a factor. This interpretation leaves the arbitrators with a wide discretion to consider those factors by themselves.

In a similar spirit, the IBA Rules, which are often applied in conjunction with the ICSID or UNCITRAL Arbitration Rules, mention in Article 9(1) that

[13] ibid.

[14] *Mr Hassan Awdi, Enterprise Business Consultants, Inc and Alfa El Corporation v Romania*, ICSID Case No ARB/10/13, Decision on the Admissibility of the Respondent's Third Objection to Jurisdiction and Admissibility of Claimant's Claims (26 July 2013) [84].

[15] Rudolf Dolzer and Christoph Schreuer, *Principles of International Investment Law* (2nd edn, OUP 2012) 285; ICSID Arbitration Rules, Articles 33–35.

[16] David Caron and Lee Caplan, *The UNCITRAL Arbitration Rules: A Commentary* (2nd edn, OUP 2013) 573.

"[t]he Arbitral Tribunal shall determine the admissibility, relevance, materiality and weight of evidence". Following the 2020 revision of the Rules, Article 9(3) was added that expressly applies to tainted evidence and reads: "[t]he Arbitral Tribunal may, at the request of a Party or on its own motion, exclude evidence obtained illegally". The precatory word "may" suggests tribunal discretion. The Prague Rules do not contain provisions on the admissibility of evidence at all, leaving it fully to the discretion of the tribunal. Although these rules are soft law, tribunals have consistently referred to them as authoritative.[17] Parties' agreements, BITs or decisions of tribunals may render these Rules binding.[18] The UNCITRAL Arbitration Rules as well as the soft law IBA Rules on the Taking of Evidence are particularly broad. The former provide no further guidance while the latter only list factors which the tribunal "shall" consider.[19] As a result, arbitrators have engaged in the exercise of balancing substantive and procedural fairness with little assistance from the arbitration rules, taking into consideration different factors in their cases.

It seems that arbitration rules do not concern the matter of admissibility — or, at the very least, do not provide obstacles or restrictions to admissibility. A more accurate statement would be to conclude that the admissibility of tainted evidence rests with the tribunals' autonomy or arbitral discretion.[20] Regardless, putting the rules to the side, it seems that the tribunals have developed their own respective criteria for admissibility within the framework of the broad arbitration rules. Accordingly, relevance,[21] credibility,[22] materiality, and also legality[23] were mentioned in the case law as separate criteria. Blair and Gojković suggest a threefold test: (a) has the evidence been obtained unlawfully by a party who seeks to benefit

[17] Cambodia Power (n 4); Hrvatska Elektroprivreda *DD* (n 4); EDF (Services) *Limited v Romania*, ICSID Case No ARB/05/13, Procedural Order No 3 (29 August 2008) [47]–[48].

[18] The IBA Arbitration Guidelines and Rules Subcommittee, *Report on the reception of the IBA Arbitration Soft Law Products* (International Bar Association 2016) 19.

[19] E.g., lack of sufficient relevance to the case or materiality to its outcome; legal impediment or privilege; unreasonable burden to produce the requested evidence; loss or destruction of the documents.

[20] Dolzer (n 15) 285.

[21] Aguas del Tunari, *SA v Republica of Bolivia*, ICSID Case No ARB/02/3, Decision on Respondent's Objections to Jurisdiction (21 October 2005) [25].

[22] ADC Affiliate *Limited and ADC & ADMC Management Limited v The Republic of Hungary*, ICSID Case No ARB/03/16, Award of the Tribunal (2 October 2006) [257]; Rumeli *Telekon AS and Telsim Mobil Telekomunikasyon Hizmetleri v Republic of Kazakhstan*, ICSID Case No ARB/05/16, Award (29 July 2008) [442]–[448].

[23] Methanex Corporation v U*nited States of America*, UNCITRAL, Final Award of the Tribunal on Jurisdiction and Merits (3 August 2005) pt II ch I [1]–[60].

from it?; (b) does public interest favour rejecting the evidence as inadmissible? And; (c) do the interests of justice favour the admission of evidence?[24]

The authors do note that no common test can be drawn for the admissibility of evidence and that the questions only serve as assistance to future tribunals.[25] Although these questions should certainly be asked by the tribunals, they are not broad enough to cover the entirety of the subject. Firstly, many acts in international law, particularly those of States, would not be deemed unlawful but rather inappropriate or unfriendly. This distinction is discussed below. Further, it is not an interest of justice that renders evidence inadmissible but rather procedural principles and various doctrines stemming from them. For these reasons this article will present the admissibility of tainted evidence as a balancing or weighing exercise between substantive fairness and procedural fairness — an approach which appears to be consistently employed by tribunals.

The difficulty is that the considerations of admissibility are scattered throughout the case law. Furthermore, no compact list of the criteria exists, which suggests that the test is not carved in stone but is flexible. This is further supported by the lack of a doctrine of precedent in investment law and the divergent views from other international and domestic courts and tribunals. To deduce a possible test for admissibility, the case law on this matter will be analysed further.

A. Weight of substance

At the outset, it is certain that illegality is not fatal to the admissibility of evidence *per se*. Whilst tribunals have frequently rejected tainted evidence, illegality was never the sole factor for such a decision. In fact, the practice of tribunals seems to be to look at the substantive value of the evidence regardless of its illegality.

Firstly, it should be noted that the practice of taking into account illegal or inappropriate evidence may not necessarily originate from arbitration or investment law, but rather from public international law. This stems from the early ICJ case of *Corfu Channel*.[26] The case concerned trespass by the British fleet in 1946 into the Corfu Channel, which was a territory claimed by Albania. The Albanian government demanded the British to obtain their consent before entry to the Channel. Prior to the trespass, and unbeknown to the British, the Channel had been mined, hence resulting in a loss of life and property to the fleet. This loss and the legality of passage over the waters triggered the dispute. The fleet entered

[24] Cherie Blair and Ema Vidak Gojković, 'WikiLeaks and Beyond: Discerning an International Standard for the Admissibility of Illegally Obtained Evidence' (2018) 33 ICSID Review: Foreign Investment Law Journal 252, 259.

[25] ibid.

[26] *Corfu Channel (UK v Albania)* (Merits) [1949] ICJ Rep 4.

the Channel on one more occasion to collect evidence. Unlike the first trespass, the second trespass was deemed outright illegal; the Court held that the United Kingdom was not allowed to collect the evidence unilaterally.[27] Interestingly, despite this holding, the ICJ then relied upon the evidence revealed in the course of the second trespass without objection from either party. One of the findings concerned the German origin of the mines which pointed at Albanian liability, given that Albania was in possession of similar mines following the Second World War.[28] In other words, although the ICJ did not make an explicit statement concerning the evidence, the Court nevertheless relied on it. Hence, the principle that all evidence can be admissible, regardless of legality or appropriateness, could originate from public international law and not arbitration. That being said, the ICJ uses neither the principle on the hierarchy of evidence nor the principle on weighing evidence. The Court's former president Judge Peter Tomka suggested the reason for this uncertainty is that the domestic principles on admissibility were never transposed into the international legal order.[29] The ICJ consequently relied solely on a broad procedural wording of the ICJ Statute in Article 48, stating that the Court "shall make all arrangements connected with the taking of evidence".

Looking at the approach taken by tribunals, one of the best examples is the *Slovenian Border Dispute*. In this inter-State Permanent Court of Arbitration case between Croatia and Slovenia, the issue was the possession of a narrow stretch of land along the two states' maritime border near the Gulf of Piran.[30] In the course of the proceedings, Croatia discovered that the Slovenian-appointed arbitrator had *ex parte* talks with one of the Slovenian counsel, discussing information about the ongoing arbitration. Such conduct pointed at the arbitrator's lack of impartiality and independence. This evidence was procured at the very least inappropriately — through the tapping of the arbitrator's phone by the Croatian intelligence.[31] Nonetheless, the arbitrator and the counsel resigned and provided apologies accordingly.

On the one hand, the *Slovenian Border Dispute* case is one of the clearest cases on the point that illegal evidence may be admissible. On the other hand, it indicates that political tensions are the supervening consequences of engaging in illegal activities, although those may be more significant in inter-State arbitrations than

[27] ibid 35.

[28] ibid.

[29] Peter Tomka and Vincent Proulx, 'The Evidentiary Practice of the World Court' in Juan Carlos Sainz-Borgo (ed), *Liber Amicorum Gudmundur Eiriksson* (San José, University for Peace Press 2016) 3.

[30] *Republic of Croatia v Republic of Slovenia*, PCA Case No 2012-04, Partial Award (30 June 2016) [80], [171], and [219].

[31] *Methanex* (n 23) pt II ch I [55].

in investment law. Tribunals hence seem to be prompted to look at the substance of the evidence as part of the balancing exercise for its admissibility.

A pivotal investment arbitration case concerning the admissibility of illegal evidence is that of *Methanex v USA*, decided under the UNCITRAL Arbitration Rules. This was a NAFTA dispute brought by a Canadian investor, alleging harm to its methanol distribution business. In the case, California imposed bans on MTBE, an additive to gasoline, because it was discovered to be a harmful carcinogen. The manufacturing of the chemical was one of the claimant investor's main activities. The claimant argued that the measure was aimed at supporting its American competitors. The tribunal, however, disagreed and found that the measure was based on legitimate scientific evidence. In this case, it was the claimant who introduced tainted evidence obtained by trespass and document theft. The documents were rejected for reasons of procedural fairness and the weight of illegality surrounding their acquisition. However, despite this finding, the tribunal did consider the question of substance of the evidence. It was stated that "the […] Documents were of only marginal evidential significance in support of Methanex's case", adding that they "could not have influenced the result of this case".[32] It is unclear if better materiality of evidence could have tilted the balance in favour of admissibility despite the extent of illegality involved. Other cases suggest it could have had this effect.

Methanex is a leading case on the admissibility of illegal and inappropriate evidence. The fact that the tribunal rendered the evidence inadmissible can create a misconception that this would be the general rule. Given the reasoning of the tribunal, however, this is not the case and the reasoning can be distinguished from other cases, both in relation to the extent of illegality (in this case there was lasting and persistent inappropriate conduct by the claimant) and in relation to the materiality of the evidence (in this case the documents had marginal relevance). *Methanex* is nonetheless authority for the proposition that substance of evidence will always be considered.

The materiality of evidence was critical in the *EDF v Romania* case. There, the British investor owned a stake in Romanian government-owned joined ventures providing airport duty-free retail services. The dispute concerned revocation of concessions given to those enterprises to provide services at several Romanian airports. The allegation here was that of inducing corruption. The claimant argued that the reason for the revocation of concessions was their failure to pay the demanded bribes, and that the revocation thus amounted to a breach of the fair and equitable treatment (FET) standard of protection in the BIT. The witness of the incident of corruption has made contradictory statements in the course of

[32] ibid [56].

proceedings. The claimant attempted to introduce new audio recordings of that witness to prove their allegation. The tribunal refused to admit such evidence on the grounds that it "lacked authenticity"[33] and that the evidence demonstrating corruption is "far from being clear and convincing" and of "doubtful value".[34] Two points are worth noting from the Award. Firstly, it confirms that the materiality of evidence is a factor in admissibility. Secondly, it mentions a requirement of authenticity. The tribunal dived deeply into the evidence's authenticity, requesting an expert opinion. The opinion reiterated that the recording was incomplete, edited, and rearranged. It was also illegal under Romanian law.[35]

Although this was a case under the ICSID Arbitration Rules, the tribunal's reasoning behind the inadmissibility was not derived from any provision found in those Rules. It could be concluded that the requirements of authenticity and materiality hence apply regardless of the applicable arbitration rules and rather stem from the arbitrators' discretion in admitting evidence. In fact, this is exactly what the tribunal agreed with when it stated:

> "[…] [s]uch discretion [to admit or reject evidence] is not absolute. In the Tribunal's judgment, there are limits to its discretion derived from principles of general application in international arbitration, whether pursuant to the Washington Convention or under other forms of international arbitration. Good faith and procedural fairness being among such principles".[36]

Good faith and procedural fairness will be discussed below. By recognising that, on the one hand, the exercise of admitting evidence is within the arbitrators' discretion. On the other hand, this discretion is not absolute and the tribunal did point at the importance of the weighing exercise. On one end of the scales lies substantive fairness; on the other end lies procedural fairness.

The *Libananco v Turkey* arbitration concerned different circumstances.[37] The dispute arose after Turkey seized electricity production and distribution companies of which the Cypriot investor Libananco possessed shares. Turkey attempted to rely on the evidence obtained by intercepting communication between the claimant's counsel and representatives. The tribunal deemed such documents to be covered

[33] *EDF* (n 12) [225].
[34] ibid [221]-[225].
[35] ibid [30]-[36].
[36] ibid [47].
[37] *Libananco Holdings Co Limited v Republic of Turkey*, ICSID Case No ARB/06/8, Decision on Preliminary Issues (23 June 2008).

by legal privilege and therefore inadmissible.[38] In particular, it was stated that "[t]he Tribunal attributes great importance to privilege and confidentiality, and if instructions have been given with the benefit of improperly obtained privileged or confidential information, severe prejudice may result".[39]

The decision points at an exception to considering substance: legal privilege. If the documents are privileged, tribunals will not consider their substantial value. The tribunal emphasised this by further adding that if a breach of confidentiality is found, "[t]hey may consider other remedies available apart from the exclusion of improperly obtained evidence or information".[40]

Finally, the *Awdi v Romania*[41] arbitration concerned a claim by an American investor holding shares in a printing company. The company held concessions from Romania to operate kiosks, which were subsequently deemed unlawful by domestic courts, therefore giving rise the claim.[42] The respondent objected to the admissibility of the claim on the grounds that the claimant was involved in actions involving human trafficking, looting of assets and businesses, crimes of running a criminal organisation, embezzlement, tax evasion, and money laundering. The evidence for the assertion was taken from ongoing, and hence confidential, Romanian domestic criminal proceedings.[43] The tribunal distinguished between the admissibility of evidence for the purpose of criminal proceedings and the probative value of the evidence for the purpose of the current arbitration:

> "[t]he issue raised by the Motion is not the admissibility of the evidence related to criminal proceedings. The issue is rather the probative value of such evidence for the purposes of this arbitration, which the tribunal is empowered to weigh and determine. In assessing this value, the tribunal shall be guided, among other things, by consideration of the presumption of innocence as a rule of public international law".[44]

In this case, the tribunal deemed the evidence to be inadmissible.[45] This suggests that when documents are obtained from ongoing domestic criminal

38 ibid [82].
39 ibid [80].
40 ibid.
41 *Mr Hassan Awdi, Enterprise Business Consultants, Inc and Alfa El Corporation v Romania*, ICSID Case No ARB/10/13, Decision on the Admissibility of the Respondent's Third Objection to Jurisdiction and Admissibility of Claimant's Claims (26 July 2013).
42 ibid [1]–[11].
43 ibid [15].
44 ibid [84].
45 ibid [1]–[11].

proceedings and are hence highly confidential, their substance will not be considered. Further, *Awdi* proves that arbitrators may be required to consider criminal concepts such as the presumption of innocence. It should be added that the tribunal in the *Awdi case* consisted of Professor Schreuer who, in his monograph cited previously, supported broad discretion of tribunals in admitting evidence by weighing the criteria of relevance, credibility, materiality, and legality.[46] This case supports his assertions, but other authorities should be referred to as well to conclude if the rules apply universally.

To conclude the point, the weighing exercise includes both substantive and procedural fairness. The materiality of evidence will always be considered with the exception of legally privileged or highly confidential documents. It is generally difficult to introduce evidence purely based on its substantive value due to procedural fairness considerations that follow. In some situations, however, the opposite is true.

B. Publicly available evidence

Having said that illegal or inappropriate evidence may, as a general rule, be admissible, there could be different reasons for this outcome. The aforementioned authorities looked at different elements of substantive fairness. There may, however, be circumstances in which procedural fairness considerations are significantly weaker and the focus of tribunals would rest on substantive fairness. If the evidence is already in the public domain, there are no interests left to be protected by the tribunals. For that reason, it is possible that the sole existence of public evidence is decisive for admissibility. Authorities suggest that this is the case. In *Gambrinus v Venezuela*, the tribunal neither considered the question of illegality nor discussed the weighing exercise.[47] The leaked Embassy Cables were quoted in the Award with no explanation as to their standing. However, given the position of public international law discussed below, outright admissibility of publicly available evidence might be inaccurately deemed a general rule as well.

In the UNCITRAL *Yukos v Russia* cases, the tribunal did rely directly on the Wikileaks evidence.[48] The string of cases, resulting in the highest investment arbitral award ever rendered at 60 billion USD, concerned the dissolution of the Russian oil company Yukos. The claims were brought by foreign shareholders alleging that the bankruptcy of Yukos was induced by the conduct of the Russian

[46] Dolzer (n 15) 285.

[47] *Gambrinus, Corp v Bolivarian Republic of Venezuela*, ICSID Case No ARB/11/31, Award of the Tribunal (15 June 2015) 44.

[48] *Hulley Enterprise Ltd v Russian Federation*, PCA Case No 2005-03/AA 226, Final Award (18 July 2014) [1218].

Federation, namely by arrests, taxation, and auctioning of assets. One of the allegations concerned the duress of Yukos' auditors, PwC, discovered in the US embassy cables by WikiLeaks. In particular, it demonstrated harassment of PwC by the Russian government in order to stop audits of Yukos and hence legitimise the latter's bankruptcy. Curiously, although neither of the parties in the case called witnesses from the company, the tribunal formed the view that the analysis of their role in the case was essential.[49] The WikiLeaks evidence was relied upon, but no view was taken on the admissibility of such evidence. The authorities quoted by the tribunal in demonstrating misconduct towards PwC were found on the WikiLeaks' website. By implication it can be concluded that such evidence can be relied upon since it was publicly available.

In the ICSID case of *ConocoPhillips v Venezuela*, one of the alleged breaches concerned Venezuela's bad faith during the negotiations between the parties about compensation for the expropriation of ConocoPhillips' assets.[50] The tribunal, however, faced an issue of a Confidentiality Agreement covering that negotiation period. Accordingly, both the claimant and the respondent were unable to provide any evidence on the matter. Venezuela, however, pointed the tribunal to the WikiLeaks US embassy cable which discussed the negotiations. It was submitted that such evidence was not covered by the Confidentiality Agreement.[51] However, the evidence was introduced at the wrong moment — after the merit phase, in the quantum phase (albeit before the Final Award). The tribunal hence rejected the respondent's Request for Reconsideration.[52]

This Decision came with a strong dissent from the arbitrator Georges Abi-Saab. He stated that failing to admit the evidence which had "a high degree of credibility"[53] constituted a "travesty of justice".[54] Not only does that statement reaffirm that inappropriate evidence should be in some circumstances admissible, it also suggests that evidence which is in the public domain cannot be omitted

[49] ibid [1184]–[1186].

[50] Con*ocoPhillips Petrozuata BV, ConocoPhillips Hamaca BV and ConocoPhillips Gulf of Paria BV v Bolivarian Republic of Venezuela*, ICSID Case No ARB/07/30, Interim Decision (17 January 2017) [70].

[51] ibid [75].

[52] Con*ocoPhillips Petrozuata BV, ConocoPhillips Hamaca BV and ConocoPhillips Gulf of Paria BV v Bolivarian Republic of Venezuela*, ICSID Case No ARB/07/30, Decision on Respondent's Request for Reconsideration (10 March 2014) [24].

[53] Con*ocoPhillips Petrozuata BV, ConocoPhillips Hamaca BV and ConocoPhillips Gulf of Paria BV v Bolivarian Republic of Venezuela*, ICSID Case No ARB/07/30, Dissenting Opinion of Georges Abi-Saab (10 March 2014) [64].

[54] ibid [67].

at any stage of the proceedings, regardless of timing requirements. In fact, as Professor Abi-Saab stated,

> "if [the Arbitrators] become aware, before the final award, that they have made a crucial error of fact or of law that led them astray in their findings, or of new evidence or changing circumstances to the same effect, they may not hesitate to revisit their decisions".[55]

If so, the dissent strongly argues why inappropriate evidence should be admitted if it is relevant. The commentators on the case agree with this view, concluding that even the majority Decision acknowledged the suitability and significance of the evidence they rejected.[56] It makes little difference that it was an ICSID case, given that the rules of the ICSID Convention were not relied upon in arriving at Professor Abi-Saab's and the commentators' conclusion. It would equally apply in UNCITRAL arbitrations.

It should be noted that both in *ConocoPhillips v Venezuela* and in *Yukos v Russia*, the tribunals did not engage deeply with the weighing of the evidence. The arguably illegal (and certainly somewhat inappropriate) evidence obtained by WikiLeaks was simply relied upon — in *ConocoPhillips*, without success due to wrong timing; in *Yukos*, directly by the tribunal and at its own initiative. Nevertheless, it cannot be correct that evidence present in the public domain outright renders other principles on admissibility redundant. There must be a limit. Otherwise, the party which has obtained the illegal evidence would simply leak it to the public domain, hence rendering it admissible. Perhaps this would point to the principles of procedural fairness, which could come into play in such circumstances. These principles will be discussed further.

Given the indeterminacy of the authorities in public international law, outright admissibility of publicly available evidence also seems incorrect. The ICJ seems not to have formed a position on the issue, despite having parties which pleaded WikiLeaks-derived evidence in certain cases. Such arguments were raised on a number of occasions in the oral hearings — for example, in *Costa Rica v*

[55] ibid [51].

[56] James Boykin and Malik Havalic, 'Fruits of the Poisonous Tree: The Admissibility of Unlawfully Obtained Evidence in International Arbitration' (2015) 5 TDM 1, 9.

Nicaragua,[57] *Macedonia v Greece*,[58] and *Croatia v Serbia.*[59] In all these cases, the ICJ did not raise the issue of the admissibility of the evidence, nor did the Court draw upon the evidence in its own judgments.[60] One possible exception could be the recent *Chagos Archipelago* case, although the ICJ merely cited the approval of the admissibility of WikiLeaks cables by the UK Supreme Court without taking any position on the matter, nor relying on the evidence.[61] In contrast, an investment tribunal may introduce publicly available evidence *sua sponte* — at its own motion — if it deems it to be relevant to the submissions which were introduced by the parties. As a consequence, it would be instantly admitted into the arbitration with the only limitation being that it must relate to the submissions which the parties have already made to afford the opponents a right to be heard.[62] There are, however, exceptions to the rule of the general admissibility of publicly available evidence.

Caratube v Kazakhstan is another Award concerning the admissibility of illegal documents, referred to as "stolen documents"[63] by the respondent. The claimants asserted that a contract for the installation and exploitation of an oil field in Kazakhstan was duly performed, with the claimants even exceeding their contractual obligations.[64] The respondent disagreed and argued that the claimants "systematically committed material breaches throughout the life of the Contract and [were] in a persistent state of material breach […] affecting virtually all areas of its activity".[65] This resulted in the revocation of the licence to exploit the oilfield,

[57] *Maritime Delimitation in the Caribbean Sea and the Pacific Ocean (Costa Rica v Nicaragua)* ICJ Verbatim Record 2017/15, 24 <www.icj-cij.org/public/files/case-related/157/157-20170713-ORA-01-00-BI.pdf> accessed 10 March 2021.

[58] *Application of the Interim Accord of 13 September 1995 (The Former Yugoslav Republic of Macedonia v Greece)* ICJ Verbatim Record 2011/6 footnotes 44 and 108 <www.icj-cij.org/public/files/case-related/142/142-20110322-ORA-01-00-BI.pdf> accessed 13 March 2021.

[59] *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Croatia v Serbia)* ICJ Verbatim Record 2014/14 [3] and [10] <www.icj-cij.org/public/files/case-related/118/118-20140311-ORA-01-00-BI.pdf> accessed 13 March 2021.

[60] Gregoire Bertrou and Sergey Alekhin, 'The Admissibility of Unlawfully Obtained Evidence in International Arbitration: Does the End Justify the Means?' (2018) 4 The Paris Journal of International Arbitration 11, 22.

[61] *Legal Consequences of the Separation of the Chagos Archipelago from Mauritius in 1965* (Advisory Opinion) [2019] ICJ Rep 95 [130].

[62] *Daimler Financial Services AG v Argentine Republic*, ICSID Case No ARB/05/1, Decision on Annulment (7 January 2015) [295].

[63] *Caratube International Oil Company LLP and Devincci Salah Hourani v Republic of Kazakhstan*, ICSID Case No ARB/13/13, Award of the Tribunal (27 September 2017) [152].

[64] ibid [38]–[50].

[65] ibid [45].

the termination of the contract, and subsequent investigations of Caratube by the governmental authorities.

The claimants made an application to the tribunal to obtain leave to introduce evidence available on the internet, such evidence being a part of around 60,000 documents leaked from the respondent's computer servers in what was known as "KazakhLeaks".[66] Although the Decision on the request is not public, the Final Award reiterates its conclusions. The tribunal did allow the claimants to produce that evidence in the arbitration. There was, however, one limitation: the tribunal explicitly protected the communications covered by the attorneyclient privilege.[67]

C. Weight of wrongfulness

In admitting evidence, the tribunals will also weigh the extent of wrongfulness. There exist, however, a number of borderline cases in international law where the scope of illegality cannot be accurately determined. Conduct that would appear criminal in domestic laws may frequently not be prohibited in international law. Three examples can be mentioned: espionage, unfriendly acts, and corruption. The more tainted the evidence, the less likely tribunals will admit the evidence into the proceedings. In addition, highly tainted evidence will prompt tribunals to ensure procedural fairness is afforded to the opponents.

A good example of the weight of illegality resulting in inadmissibility of tainted evidence is the *Methanex* case.[68] The illegal activities included "deliberate trespass onto private property and rummaging through dumpsters inside the office-building for other persons' documentation".[69] Although the conduct was not criminal under Californian law, it was a civil breach. The tribunal ruled against the admissibility of the evidence, basing its decision on the principles of good faith, justice, and fairness.[70] The tribunal's reasoning seems to centre around the scope and extent of illegal activities; the outcome of the case can hence be isolated to the particular facts. Concerning the tainted evidence, it was stated that "this documentation was obtained by successive and multiple acts of trespass committed by Methanex over five and a half months in order to obtain an unfair advantage over the USA as a Disputing Party to these pending arbitration proceedings".[71]

66 ibid [150].
67 ibid [156].
68 *Methanex Corporation v United States of America*, UNCITRAL, Final Award of the Tribunal on Jurisdiction and Merits (3 August 2005).
69 ibid [55].
70 ibid [60].
71 ibid [59].

In the case, the extent of illegality was simply so high that it would outweigh any substantive value of the evidence. The tribunal further noticed that the conduct took place both before and after the arbitration was constituted. Presumably, in cases where the breaches were not as heinous and persistent with an obvious objective to influence the arbitration, the evidence could indeed be admitted.

Moving on to espionage, it is a frequent method of obtaining evidence by States. The investment case of *Libonanco v Turkey* explicitly labelled "surveillance and interception of communications" as amounting to possible espionage.[72] Obtaining documents through espionage was also exactly what happened in the ICJ case of *Tehran Hostages*.[73] Iran obtained a number of confidential USA Embassy documents in the course of seizure of its premises.[74] Despite this wrongful conduct, the Court did not condemn espionage committed by Iran to be an illegal act. The possible explanation could be that of Professor Schaller, who wrote:

> "[e]spionage is regarded by States as a necessary tool for pursuing their foreign policy and security interests, and for maintaining the balance of power at the inter-State level […]. Accordingly, there is no general prohibition of espionage in international law, and it is unlikely that such a prohibition will emerge in the future".[75]

In accordance with this statement, espionage is inherently a tool of States, not private entities. For similar reasons, while espionage is a criminal offence under domestic laws, it would not be in public international law. This is not, however, equivalent to saying that the tribunals would disregard the use of espionage altogether. It could still be deemed an unfriendly act, as will be discussed below, and hence it would still be wrongful and proof of impropriety.[76] In other words, it would be relevant for the balancing exercise, although carrying a smaller weight for the tribunal.

It is clear that there is no customary rule of friendship between States in international law. Therefore, acts of States cannot be deemed unfriendly without

[72] *Libananco Holdings Co Limited v Republic of Turkey*, ICSID Case No ARB/06/8, Excerpts of Decision on Annulment (22 May 2013) [170].

[73] *United States Diplomatic and Consular Staff in Tehran (United States of America v Iran)* (Judgment of 24 May 1980) [1980] ICJ Rep 3.

[74] ibid [82].

[75] Christian Schaller, 'Spies', *Max Planck Encyclopedia of Public International Law* (2015) <https://opil.ouplaw.com/view/10.1093/law:epil/9780199231690/law-9780199231690-e295> accessed 16 December 2020.

[76] *Military and Paramilitary Activities in and against Nicaragua Case (Nicaragua v United States of America)* (Merits) [1986] ICJ Rep 14, [272]-[274].

a legal basis. The ICJ in the *Military and Paramilitary Activities in and against Nicaragua* case stated:

> "[s]uch a duty might of course be expressly stipulated in a treaty, or might even emerge as a necessary implication from the text; but as a matter of customary international law, it is not clear that the existence of such a far-reaching rule is evidenced in the practice of States".[77]

Most investment treaties, however, would contain a fair and equitable standard of treatment of investors.[78] Investment tribunals would take unfriendly acts into consideration both as a part of FET violations,[79] as well as independently, even if the conduct is not a breach of the investment treaty.[80] In short, such conduct would be considered by tribunals in assessing the admissibility of evidence, despite not being wrongful under public international law.

It was previously said that corruption is illegal in most, if not all, legal systems.[81] Equally, it can be a breach of international public policy. However, such a breach is not illegal *per se*, although the tribunal in *Hamester v Ghana* did state that an investment created by means of corruption will lose protection.[82] There are neither cases nor doctrine addressing evidence that was obtained directly through corruption. However, following from the aforementioned case, such evidence would be tainted with a high degree of impropriety for the sake of balancing its admissibility. Given that corruption may refute the protection of an investment altogether in investment law, it is likely that evidence procured though corruption would be outright inadmissible — something that would never be the case for other

[77] ibid [273].

[78] More than 2000 BITs contain the FET standard. See UNCTAD, 'World Investment Report 2002' (*UNCTAD*, 12 June 2003) <https://unctad.org/system/files/official-document/wir2002_en.pdf> accessed 18 January 2020.

[79] *Mondev International Ltd v United States of America*, [2002] ICSID Case No ARB(AF)/99/2, Award [118]–[119].

[80] MCI Power Group *LC and New Turbine, Inc v Republic of Ecuador*, ICSID Case No ARB/03/6, Award (31 July 2007) [371].

[81] *World Duty Free* (n 8).

[82] *Gustav F W Hamester GmbH & Co KG v Republic of Ghana*, ICSID Case No ARB/07/24, Award (18 June 2010) [123]-[124].

types of illegality and impropriety. Such a strict approach is consistent with some, but not all, domestic laws.

D. Position under comparative law

Domestic procedural legal systems are inconsistent concerning the issue of admissibility of illegal or inappropriate evidence. Some jurisdictions seem to be liberal, while others appear to be strict. The so called 'unified legal system' doctrine states that domestic substantive law may not be inconsistent with procedural law. If a jurisdiction did employ this doctrine, illegally obtained evidence would be frequently deemed inadmissible due to its illegality. The converse principle is that of the 'theory of segregation', whereby substantive and procedural laws are distinguished and separated.[83] Again, different jurisdictions take divergent views on whether and to what extent such principles are applicable. It should be noted, however, that the discussion is moot in arbitration where most of the procedural aspects of conducting the arbitration (and hence admissibility) are in the exclusive competence of the arbitrators.

Most European legal systems do not explicitly regulate the handling of illegally obtained evidence.[84] This appears to follow the European Court of Human Rights' (ECtHR) judgment in the *Schenk v Switzerland* case, where the ECtHR stated that there is no general prevention on the admissibility of such evidence under the European Convention on Human Rights (ECHR), although in specific circumstances it may breach the Convention rights.[85] Under English law, for instance, there is also no provision that excludes the admissibility of illegally obtained evidence. It seems that English judges would be more concerned with the materiality of such evidence, much like arbitral tribunals, although their tolerance in doing so is not as high. The major factor for the weighing exercise in English law seems to be public policy.

The British Human Rights Act 1998 transposes certain articles of the ECHR into domestic law. Tapping phones and hacking communications are listed as possible interferences with the rights pursuant to it.[86] This approach was taken

[83] Bettina Nunner-Krautgasser and Philipp Anzenberger, 'Inadmissible Evidence: Illegally Obtained Evidence and the Limits of the Judicial Establishment of the Truth' in Vesna Rijavec, Tomaž Keresteš and Tjaša Ivanc (eds) *Dimensions of Evidence in European Civil Procedure* (Kluwer Law International 2016) 198.

[84] ibid.

[85] *Schenk v Switzerland* App no 10862/84 (ECtHR, July 12 1988) [46].

[86] Anupreet Amole and Jane Colston, 'Fruit from A Poisoned Tree: Unlawfully Obtained Evidence' (*The Law Society Gazette*, 30 August 2017) <www.lawgazette.co.uk/commentary-and-opinion/fruit-from-a-poisoned-tree-unlawfully-obtained-evidence/5062566.article> accessed 22 August 2019.

in the case of *Jones v University of Warwick*, in which a breach of the right to privacy was found and the evidence obtained using a hidden camera was not admitted by the court.[87] Lord Woolf CJ, giving the judgment, acknowledged that the test was that of reconciling "conflicting public policies" which accordingly have to be balanced against each other.[88] In doing so, he stated that the leading notion is that of achieving justice.[89] This principle can be, however, restricted to the circumstances, given that the case concerned insurance entitlement and the illegal evidence was aimed at refusing the claimant's right to such insurance.[90] It hence carried a strong public policy implication that could have been the reason for the refusal of its admissibility. In *Rall v Hume*, a personal injury case also concerning video evidence, the court did admit the inappropriate evidence and allow the defendant to confront the claimant with such evidence in cross-examination.[91] Personal injury is still a public policy-heavy area. It can therefore be concluded that in purely commercial cases, the materiality of the evidence would be weighed and likely admitted by the court due to fewer public policy considerations. Such considerations, however, add to legal uncertainty. Burrough J in *Richardson v Mellish* contended that public policy is "a very unruly horse, and once you get astride of it you never know where it will carry you".[92] Lord Denning, however, later responded that "with a good man in the saddle, the unruly horse can be kept in control. It can jump over obstacles".[93]

A possibly common approach between the English courts and the ICJ concerns the admissibility of publicly available evidence. The British Supreme Court recently admitted leaked diplomatic cables in a case concerning the challenge to the UK government's handling of decolonisation of the Chagos Archipelago.[94] The suit was brought by native Chagossians who were by English law prohibited from returning to their ancestral homes on the islands. This approach was quoted a year later, but not discussed, on appeal to the ICJ in the Advisory Opinion.[95]

Turning to civilian jurisdictions, the Austrian Civil Procedure Code does not contain provisions regulating the admissibility of illegal evidence. The general

87 *Jones v University of Warwick*, [2003] EWCA Civ 151.
88 ibid [21].
89 ibid.
90 ibid [23].
91 *Rall v Hume* [2001] 3 All ER 248, 254.
92 Richardson v M*ellish* [1824] 2 Bing 229, 252 (Borrough J).
93 *Enderby Town Football Club Ltd v The Football Association Ltd* [1971] Ch 591, 606 (Denning MR).
94 *R (on the application of Bancoult No 3) v Secretary of State for Foreign and Commonwealth Affairs* [2018] UKSC 3 [9].
95 *Chagos Archipelago* (n 61). The ICJ misquoted the British Supreme Court by stating that evidence in the diplomatic cables was held by them to be inadmissible. In fact, the opposite was true. There is hence a mistake in the Advisory Opinion although this does not impact the conclusions of the ICJ.

position is that such evidence should be considered.[96] Austria therefore leans towards the aforementioned theory of segregation, whereby procedural law is separated from substantive law, although no such definitive statements were made by the courts or commentators. This seems to be particularly the case for civil proceedings — although in other areas, such as under data protection law, the outcome of the analysis would be different.[97] In the latter circumstances, the courts are in fact inclined to take a similar policy-balancing exercise as English courts do and take into account the adverse party's right of personality and right to data protection. When it comes to criminal proceedings, Austria also does not deem illegal evidence to be outright inadmissible. Quite to the contrary, consideration of all evidence is an obligation imposed on the courts as part of the principles of truth-finding and freedom of evidence.[98] The EU Commission Panel, monitoring Member States compliance with fundamental rights, summarised that Austria does not know the American rule of the "fruit of the poisonous tree".[99]

In the USA, contrary to the position in common law jurisdictions which follow the English model, questions on the admissibility of illegal and inappropriate evidence do not exist. Illegally obtained evidence, e.g., evidence procured through a criminal act, such as corruption or fraud, would not be legal in itself and would not be admissible. In the USA, the doctrine of the fruit of the poisonous tree is applied, stating that the manner in which evidence was acquired ('the tree') taints the evidence ('the fruit'), and that this, in turn, will render the evidence inadmissible.[100]

The doctrine is by no means limited to the USA. It possibly exists in ECtHR's jurisprudence as formulated in the *Gäfgen v Germany* case,[101] although it contrasts with the earlier judgment of *Schenk v Switzerland*.[102] Turning to investment law, some commentators also argue that the fruit of the poisonous tree principle could have been the reason for the decision on the inadmissibility of evidence in

[96] Patrick Mittlboeck, 'Austria: Use Of Unlawfully Obtained Evidence In Austrian Civil Proceedings?' (*Mondaq*, 5 April 2019) <www.mondaq.com/Austria/x/796414/Civil+Law/Use+Of+Unlawfully+Obtained+Evidence+In+Austrian+Civil+Proceedings> accessed 26 August 2019.

[97] 6 Ob 16/18y (The Austrian Supreme Court of Justice 2018).

[98] 15 Os 3/92-8 (The Austrian Supreme Court of Justice 2018).

[99] EU Network of Independent Experts on Fundamental Rights, 'Opinion on the status of illegally obtained evidence in criminal procedures in the Member States of the European Union' (*EU Commission*, 30 November 2003) <https://sites.uclouvain.be/cridho/documents/Avis.CFR-CDF/Avis2003/CFR-CDF.opinion3-2003.pdf> accessed 26 August 2019.

[100] *Silverthorne Lumber Co v United States* 251 US 385, 385 (1920).

[101] *Gäfgen v Germany* App no 22978/05 (ECtHR, 18 March 2009) [29] (evidence was obtained in breach of another Convention right).

[102] *Schenk v Switzerland* App no 10862/84 (ECtHR, 7 December 1988) [46] (the case concerned pre-internet evidence so may be distinguished from circumstances in more recent cases cited).

the *Methanex v USA* case. Both parties quoted American law extensively and the tribunal appears to have contended that the actions of the claimant were illegal as a matter of United States law.[103] The fruit of the poisonous tree could therefore well become a general principle of law recognised by States, pursuant to Article 38(1) of the ICJ Statute — although as was discussed above, it is by no means a ubiquitous principle across jurisdictions.

To conclude the point, the approaches of different jurisdictions and legal systems are inconsistent. They form a patchwork of practices and tests that provide little indication about any notion of a general principle of law.

IV. Procedural Principles

In light of the discussion above, it is clear that parties against whom tainted evidence is used may be at a significant disadvantage. Since materiality is relevant and can trump the means of obtaining evidence, stronger parties in investment cases can utilise their vast resources to obtain favourable evidence and conceal unfavourable evidence. Furthermore, although States and investors are considered equal parties once an investment case has been brought,[104] they are inherently different entities. On the one hand, a powerful superstate has greater resources at their disposal than a private investor. On the other hand, an international corporation may be more powerful in the proceedings than a small, less economically developed State. Further, as was mentioned, States have a variety of international practices which would be unavailable or illegal for the investors, particularly in the field of espionage and other domestically criminal activity. The scope of international legal personality (and hence the capacity to possess rights and obligations in international law) as well as the capacity to act in international law (which presupposes legal personality and includes the standing to bring a claim in international law) are different for States and investors.[105] Individual subjects of international law differ in the nature, extent, or existence of their rights.[106]

This is where principles of procedural fairness, good faith, and clean hands come into play—they ensure, to an extent, that both parties are on a level playing field. Procedural fairness in particular contains the principle of equality of arms, which would be triggered in circumstances of attempts to introduce illegal evidence.

[103] Bertrou (n 60).

[104] See the Convention on the Settlement of Investment Disputes between States and Nationals of Other States (opened for signature 18 March 1965, entered into force 14 October 1966) ('ICSID Convention'), Article 25.

[105] Malcolm Shaw, *International Law* (8th edn, CUP 2017) 166–170.

[106] *Reparation for Injuries Suffered in the Service of the United Nations* (Advisory Opinion) [1949] ICJ Rep 174, 178; *LaGrand (Germany v United States of America)* (Judgment of 27 June 2001) [2001] ICJ Rep 466, 494.

Since such actions include one party taking inappropriate measures against the other, considerations of good faith and clean hands are simultaneously triggered. These principles will be analysed in turn, but the starting point is the weight of wrongfulness that tribunals attach to the evidence in determining its admissibility.

A. Procedural Fairness and Equality of Arms Doctrines

Article 15(1) UNCITRAL Arbitration Rules contains the requirement of procedural fairness:

> "[s]ubject to these Rules, the arbitral tribunal may conduct the arbitration in such manner as it considers appropriate, provided that the parties are treated with equality and that at any stage of the proceedings each party is given a full opportunity of presenting his case".

The tribunal in *Methanex* was clear that equality of arms is both required pursuant to the above procedural provision, and also as a "general legal duty" owed by the disputing parties to one another and to the tribunal.[107]

The ICSID Convention contains a similar requirement. Article 52 states that the breach of a fundamental rule of procedure forms one of the grounds for a request for annulment of the award:

> "(1) Either party may request annulment of the award by an application in writing addressed to the Secretary-General on one or more of the following grounds: […] (d) that there has been a serious departure from a fundamental rule of procedure".

The 1958 New York Convention contains similar grounds for reusal of recognition and enforcement of awards.[108] It was clarified in *Wena Hotels v Egypt* that a departure is "serious" where it is "substantial and [is] such as to deprive a party of the benefit or protection which the rule was intended to provide".[109] Equally, "fundamental" in the above provision refers to the "set of minimal standards of procedure to be respected as a matter of international law".[110] Even marginal departure from procedural standards can therefore subject an award to

[107] *Methanex* (n 23) pt II ch I [54].

[108] 1958 New York Convention on the Recognition and Enforcement of Foreign Arbitral Awards, Article V(1)(d).

[109] Wena Hotels L*td v Arab Republic of Egypt*, ICSID Case No ARB/98/4, Decision (Annulment Proceeding) (5 February 2002) [58].

[110] ibid [57].

annulment.[111] In *Giovanni Alemanni v Argentina*, the tribunal not only stated that a "fundamental rule of procedure" in the provision includes the equality of arms, but also noted that the principle would apply even in the absence of Article 52 since it is "fundamental to the judicial process".[112] The *Giovanni Alemmani* case instead concerned due process,[113] but the tribunal nonetheless deemed it necessary to reiterate equality of arms as well. This only demonstrates the fundamental importance of the doctrine.

In investment arbitration, contrary to commercial arbitration, the parties to proceedings are significantly different from one another. The respondent is always a sovereign State. States have at their disposal resources which claimants would not have. This includes capital, intelligence, or different rights in international law. Investors can be individuals — or, more frequently, multinational corporations with revenue greater that some of the world's States. Given that illegally and inappropriately obtained evidence is not automatically inadmissible, an argument could be made that investment law encourages resorting to illegitimate means to obtain such evidence to the disadvantage of the opposing party. In *Libananco v Turkey*, the State engaged in "surveillance and interception of communications"[114] of the claimant. This gave the respondent State access to "hundreds, or even thousands, of counsel's communications with their clients",[115] which the State then tried to use as evidence. This feat could not have been achieved without the resources at the State's disposal. The tribunal noted that admitting such evidence would cause "irrevocable prejudice to [claimant's] position in this arbitration".[116]

The parties against whom tainted evidence is used, however, are not unprotected against such misconduct. The principle of equality of arms would be

[111] Further examples: Klöckner Industrie-Anlagen GmbH and others v *United Republic of Cameroon and Société Camerounaise des Engrais*, ICSID Case No ARB/81/2, Decisions of the Ad Hoc Committee (Unofficial English Translation) (3 May 1985) [82]–[113]; CDC Group plc v Republic of Seychelles, ICSID Case No ARB/02/14, Decision on Annulment (29 June 2005) [48]–[49]; Azurix Corporation v *The Argentine Republic*, ICSID Case No ARB/01/12, Decision on the Application for Annulment of the Argentine Republic (1 September 2009) [49]–[52] and [234]; Enron Corporation and Ponderosa Assets, *LP v The Argentine Republic*, ICSID Case No ARB/01/3, Decision on the Application for Annulment of the Argentine Republic (30 July 2010) [70]–[71].

[112] Giovanni Alemanni and Others v *The Argentine Republic*, ICSID Case No ARB/07/8, Decision on Jurisdiction and Admissibility (17 November 2014) [323].

[113] ibid [321]–[325].

[114] *Libananco* (n 72).

[115] ibid [72].

[116] ibid.

applied by tribunals to recognise the imbalance between parties. Article 9(2)(g) of the IBA Rules on the Taking of Evidence reads:

> "The Arbitral Tribunal shall, at the request of a Party or on its own motion, exclude from evidence or production any Document, statement, oral testimony or inspection for any of the following reasons: […] (g) considerations of procedural economy, proportionality, fairness or equality of the Parties that the Arbitral Tribunal determines to be compelling".

In a Commentary to the above Rules, Article 9(2)(g) is labelled as a "catch-all" provision.[117] One example given is that it would apply in situations of inconsistencies between jurisdictions concerning privileged documents. One party cannot take advantage of softer laws on document privilege in one jurisdiction.[118] Undoubtedly, the same reasoning would apply to a State using its own inherent resources which are unavailable to the other party to produce evidence. Article 9(2)(g) is intended to "help ensure the arbitral tribunal provides the parties with a fair, as well as an effective and efficient, hearing".[119] Tribunals should be particularly alert to the consequences of illegality of evidence given that, under the recently-added Article 9(3) in the 2020 IBA Rules, they may reject such evidence.

The Prague Rules on the Efficient Conduct of Proceedings in International Arbitration — a civil law competitor of the IBA Rules — have a similar thrust.[120] Article 1(4) of the Prague Rules states that "[a]t all stages of the arbitration and in implementing the Prague Rules, the arbitral tribunal shall ensure fair and equal treatment of the parties and provide them with a reasonable opportunity to present their respective cases". Although the Prague Rules do not apply procedural fairness as clearly to the admissibility of evidence as the IBA Rules, the application of "fair and equal treatment of the parties" during "all stages of the arbitration" would carry a similar result.

Such an approach to admissibility in procedural rules is a good starting point to explain the decisions of tribunals to admit or refuse evidence. The aforementioned *Slovenian Border Dispute* case can hence be easily distinguished.

[117] 1999 IBA Working Party and 2010 IBA Rules of Evidence Review Subcommittee, 'Commentary on the Revised Text of the 2010 IBA Rules on the Taking of Evidence in International Arbitration', (2011) 5(1) Dispute Resolution International 45, 77–78.

[118] ibid.

[119] ibid.

[120] Sol Argerich, 'A Comparison of the IBA and Prague Rules: Comparing Two of the Same' (*Kluwer Arbitration Blog*, 2 March 2019) <http://arbitrationblog.kluwerarbitration.com/2019/03/02/a-comparison-of-the-iba-and-prague-rules-comparing-two-of-the-same/> accessed 12 December 2019.

There, the evidence was admitted because the issues of equality of arms would never arise; it was an inter-State arbitration between parties of comparable wealth (Croatia and Slovenia). In such circumstances, it would be more demanding to demonstrate that the admission of Slovenian evidence breached equality of arms when Croatia had the same tools at their disposal.

Equality of arms arguments were developed more deeply in investment cases. In the aforementioned *Methanex v USA* arbitration, the respondent argued the converse — that good faith should prevent the claimant from having its evidence admitted since it was obtained in the course of burglaries. The tribunal agreed and quoted the principle of equality of arms.[121] The principle of equality of arms hence protects both States and investors. In fact, the tribunal clearly stated that "just as it would be wrong for the USA *ex hypothesi* to misuse its intelligence assets to spy on Methanex (and its witnesses) and to introduce into evidence the resulting materials into this arbitration, so too would it be wrong for Methanex to introduce evidential materials obtained by Methanex unlawfully".[122]

A similar conclusion was reached in the *Caratube II v Kazakhstan* arbitration.[123] In attempting to convince the tribunal that illegal, publicly available evidence is not admissible, the respondent used the argument that they did not have access to the claimant's emails.[124] The need to preserve the truthfulness of the award was deemed to outweigh the potential unfairness that might have resulted in admission.[125] When protecting State parties, slightly different considerations would apply. The *Caratube* tribunal noted explicitly that the fact that the respondent is a State is relevant and that tribunals must "be mindful when issuing provisional measures not to unduly encroach on the State's sovereignty and activities serving public interests".[126] Needless to say, a request for provisional measures can include decisions on admissibility and hence the application of States' interests in preserving sovereignty and public interests is an overarching aim for arbitral

[121] *Methanex* (n 23) pt II ch I [1] and [53].
[122] ibid [54].
[123] Decision not public but was reported in secondary sources: see Bertrou (n 60).
[124] ibid.
[125] ibid.
[126] *Caratube International Oil Company LLP and Devincci Salah Hourani v Republic of Kazakhstan*, ICSID Case No ARB/13/13, Decision on the Claimants Request for Provision Measures (4 December 2014) [121].

tribunals. This view seems to be supported by the tribunals' and domestic courts' consistent practice of considering State sovereignty in treaty interpretation.[127]

There is, however, a significant exception to the principle of equality of arms. In the *Daimler v Argentina* case, the tribunal stated that once the parties received the opportunity to make submissions, the tribunal could, *sua sponte*, introduce evidence that is in the public domain.[128] The arbitrators clarified that such an exercise of their authority would not violate any principle of due process,[129] which encompasses the equality of arms. This is an important exception because it outlines the limits of equality of arms — it relates only to the 'combatants' who raise their arms against one another and not the arbitrators themselves *vis-a-vis* the parties. Given that a tribunal ultimately renders an award binding on both parties, there can be no equality between the two.

B. Good faith and clean hands doctrines

Good faith is a general principle of law.[130] The clean hands doctrine is arguably a general principle of law as well, although recent authorities speak against its existence.[131] However, the two will be discussed in parallel, given that, in relation to the admissibility of tainted evidence, a breach of one of these principles would frequently be a breach of the other, and arbitral tribunals have often not distinguished them.

The clean hands doctrine would be applicable in considerations of admissibility, both of the entire claims and of evidence.[132] In the *Factory at Chorzów* case, the Permanent Court of International Justice stated that clean hands is a "principle generally accepted […] if the former party has by some illegal act

[127] See El Paso Energy International Company v *The Argentine Republic*, ICSID Case No ARB/03/15, Decision on Jurisdiction (27 April 2006) [70]; Pan American Energy LLC and BP Argentina Exploration Company v *The Argentine Republic*, ICSID Case No ARB/03/13 and BP America Production Company, Pan American Sur SRL, Pan American Fueguina, SRL and Pan American Continental SRL v Argentine Republic, ICSID Case No ARB/04/8, Decision on Preliminary Objections (27 July 2006) [99]; Sanum Investments Limited v Lao People's Democratic Republic PCA Case No 2013-13, Judgment of Singapore High Court (20 January 2015) [124].

[128] *Daimler Financial Services* (n 62).

[129] ibid.

[130] *Nuclear Tests (Australia v France)* (Judgment of 20 December 1974) [1974] ICJ Rep 253, 253 and 267.

[131] South American Silver Limited v Bolivia, PCA Case No 2013-15, Award (30 August 2018) [436]–[453]; Hesham Talaat M A*l-Warraq v The Republic of Indonesia*, UNCITRAL, Final Award (15 December 2014) [646].

[132] See *Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory* (Advisory Opinion) [2004] ICJ Rep 136, 163. It was one of the arguments raised by Israel. The ICJ, however, did not rely on the principle in their Opinion, nor disagreed with its application. The case nonetheless illustrates that the argument of clean hands is raised at the stage of admissibility.

prevented the latter from fulfilling its obligation in question".[133] This seems to be a rather restricted and old statement of the Court and may no longer apply.

In *Methanex*, the tainted evidence also was rejected for breaches of good faith and the clean hands doctrine, along with a breach of equality of arms.[134] Therefore, such considerations appear to be highly relevant to the admissibility of evidence. It is supported by the statement in *Awdi v Switzerland*, which deemed public international law to be part of the weighing exercise of admissibility of tainted evidence.[135] Given that the good faith and clean hands doctrines stem from public international law, they would fall under the weighing exercise considerations on the admissibility of tainted evidence. On the other hand, in the *Slovenian Border Dispute*, the tribunal did not proceed to consider any of these principles.[136]

Based on *Methanex*, *Awdi* and the *Slovenian Border Dispute*, the relationship between procedural fairness, good faith, and clean hands can be established. Tribunals would first look at procedural fairness. This practice is consistent with the approach of public international law to general principles of law. General principles of law are subsidiary authorities and will be considered in the ICJ jurisprudence "for filing a gap in the treaty or customary rules available to settle a particular dispute and […] will decline to invoke them when such other rules exist".[137] It was previously discussed that procedural fairness stems from arbitration rules and hence treaties. In *Methanex,* it was not conclusive for the tribunal to have the evidence rejected based on equality of arms alone. The arbitrators hence also added the principles of unclean hands and good faith to their reasoning.

V. Tools for Solving the Imbalance

The above sections discussed the balancing exercise. In some cases, however, the imbalance that would lead to the inadmissibility of evidence might be addressed by the arbitrators themselves using the tools available to them, that is, judicial assistance and production of document orders. Exercising such powers

[133] Case concerning the Factory at *Chorzów* (Judgment of 26 July 1927) PCIJ Series A No 9 31.
[134] *Methanex* (n 23) pt II ch I [1] and [53].
[135] *Mr Hassan Awdi* (n 14) [1]–[11].
[136] *Croatia* (n 30).
[137] Bruno Simma and others (eds), *The Charter of the United Nations: A Commentary* (3rd edn, OUP 2012) 780; *Right of Passage over Indian Territory (Portugal v India)* (Merits) [1960] ICJ Rep 6, 43.

might assist the tribunals in ensuring a just outcome of the case without striking out relevant information.

A. Judicial assistance

Arbitral tribunals are inherently ill-equipped for dealing with matters of illegality in proceedings. Not only is the standard of proof high for such issues, but arbitrators also lack the coercive powers that domestic courts would have when facing criminal charges.

Emanuele pointed that tribunals lack a number of competences:

> Power to order the production of evidence in possession, custody or control of a person who is not a party to the arbitration;
>
> Powers to impose criminal sanctions against parties who fail to comply with document production and evidentiary orders and;
>
> Powers to compel the attendance of witnesses under the penalty of fines or imprisonment and under oath.[138]

As was demonstrated by the aforementioned cases, the introduction of inappropriate evidence frequently prompts the tribunals to consider criminal or allegedly illegal behaviour. Hence, it becomes apparent that the coercive powers of the tribunals are deficient. A solution to this issue is judicial assistance. A tribunal, on its own motion or at the request of a party, would petition a domestic court to use its coercive powers in evidence production. It is therefore a method that arbitrators may use to preserve the procedural fairness between the parties to the dispute. It could be particularly useful where the inappropriate evidence is materially relevant, but its admissibility would breach equality of arms. In such circumstances, the tribunal could seek judicial assistance in finding counterarguments for the other party.

Enforcing criminal laws is ultimately the competence of national courts. It is hence in circumstances when such issues arise that the courts should be most willing to grant assistance. Judicial assistance is not, however, limited to assistance with criminal findings.

The source of competences to request assistance rests with the applicable *lex arbitri*. It is, for instance, permissible under the Swiss Private International Law

[138] Ferdinando Emanuele and others, 'State Court Assistance in the Taking of Evidence' in Ferdinando Emanuel and others (eds) *Evidence in International Arbitration: The Italian Perspective and Beyond* (Thomson Reuters 2016) 138–139.

Act (Article 184) as well as the English Arbitration Act (Section 43). The latter states:

> "[a] party to arbitral proceedings may use the same court procedures as are available in relation to legal proceedings to secure the attendance before the tribunal of a witness in order to give oral testimony or to produce documents or other material evidence".

It should be noted that granting judicial assistance is ultimately a question of domestic law. This was well demonstrated in the recent English case of *A and B v C, D and E* where the Commercial Court refused to compel a UK-based third party to submit evidence on the basis that it was not party to the arbitration agreement in a New York-seated arbitration. Consequently, the Commercial Court held that it lacked jurisdiction to compel third parties to give evidence.[139] At the same time, it was noted that the decision would be different under the laws of Hong Kong,[140] suggesting that the English approach is not shared by other common law jurisdictions.

Further, tribunals should take care not to exceed its powers by permitting such a request for evidence, which could be deemed to be *ultra petita* — that is, exceeding what the parties requested of the tribunal. Binder concluded that the tribunals should "act with great delicacy" when exercising this competence.[141]

B. Production of documents

Another method of preserving fairness between parties is prompting document production. It could be useful because it would preserve the materiality of the illegal evidence, hence allowing the tribunals to avoid refusing admissibility of potentially relevant materials while protecting the equality of arms between parties. This can be achieved in two ways. Firstly, it may be used to prompt the party which did not introduce illegal evidence to bring forward its own counterevidence on the matter. In other words, a tribunal may assist the party in providing both perspectives on an issue. However, this would be redundant in most cases, given that a party would have provided such evidence on its own motion and would not need encouragement from the tribunal. That being said, it may prompt the party to search for evidence in sources which they did not previously consider, such as

[139] A *and B v C, D and E* [2020] EWHC 258 (Comm) [32]–[33].
[140] ibid [18].
[141] Peter Binder, *International Commercial Arbitration and Mediation in UNCITRAL Model Law Jurisdictions* (4th edn, Kluwer Law International 2019) 388–396.

in the public domain. The second reason for document production is to prompt the party introducing the illegal evidence to provide further evidence. This may be particularly useful in the weighing exercise. Further evidence will provide a more complete picture of the newly introduced documents. In particular, tribunals may request an explanation of how the illegal evidence was obtained.

Using the tool of document production as mentioned is warranted by the arbitration rules. Under the ICSID and UNCITRAL Arbitration Rules, it is a basic competence of tribunals. The ICSID Rules state in Article 34:

> "(2) The Tribunal may, if it deems it necessary at any stage of the proceeding:
> (a) call upon the parties to produce documents, witnesses and experts; and
> (b) visit any place connected with the dispute or conduct inquiries there.
> (3) The parties shall cooperate with the Tribunal in the production of the evidence and in the other measures provided for in paragraph (2). The Tribunal shall take formal note of the failure of a party to comply with its obligations under this paragraph and of any reasons given for such failure".

The proposition that tribunals may call for production of documents in order to ensure equality between parties is supported by the passage quoted above. A number of observations can be made.

Firstly, the tribunal may only exercise the power to request document production under Article 34 if it "deems it necessary". Ensuring equality of arms would certainly be such a situation. If equality of arms was not observed, the entire award could be subject to annulment as was discussed in *Libananco v Turkey*.[142] It should be noted, however, that the tribunal being merely selective of the evidence provided — even disregarding some evidence entirely — should not immediately give rise to a situation of necessity. Tribunals also fulfil a "judicial function of choosing which evidence it finds relevant and which it does not".[143] But the considerations of equality of arms, even given arbitrators' wide discretion in admissibility of evidence, would prompt the tribunal to consider the questions. After all, a tribunal has a primary obligation to ensure the enforceability of the

[142] *Libananco* (n 72) [226] (applied ICSID Arbitration Rules (2006)).

[143] *Tulip Real Estate and Development Netherlands BV v Republic of Turkey*, ICSID Case No ARB/11/28, Decision on Annulment (30 December 2015) [149].

award — or, more precisely, to ensure that the award is not outright susceptible to a challenge.[144]

Secondly, the ICSID Rules explicitly mention that document production requests can be given "at any stage of proceedings". As it was discussed previously, the timing of the introduction of the tainted evidence is critical for its admissibility. Tribunals could, instead of rejecting these documents, prompt the other party to introduce its own counterevidence or request clarifications.

Further, there is a wide variety of evidence that tribunals can order. It would not be limited to written evidence and would even cover site visits. The possibility to call witnesses and conduct site visits would be particularly useful for retaining equality of arms since it introduces an objective perspective into the evidence. This in turn assists tribunals in the weighing exercise by making the assessment fairer.

Finally, on this point, paragraph 3 of Article 34 of the aforementioned provision gives arbitrators an extent of coercive powers in document production. Not only does it stipulate that the "parties shall cooperate" with tribunals in exercising this power but also that a "formal note" will be taken in cases of lack of cooperation. A failure to comply with a document production order will in and of itself constitute evidence.

Turning to the UNCITRAL Arbitration Rules, Article 24 states the following:

> "2. The arbitral tribunal may, if it considers it appropriate, require a party to deliver to the tribunal and to the other party, within such a period of time as the arbitral tribunal shall decide, a summary of the documents and other evidence which that party intends to present in support of the facts in issue set out in his statement of claim or statement of defence.
>
> 3. At any time during the arbitral proceedings the arbitral tribunal may require the parties to produce documents, exhibits or other evidence within such a period of time as the tribunal shall determine".

Contrary to the ICSID Arbitration Rules, the UNCITRAL Rules do not explicitly confer on tribunals coercive powers in ordering document production. Instead, they allow tribunals to first request a period of notice before new evidence

[144] Considering, for example, the decision in Achmea in Eskosol SpA in liquidazione v Italian Republic, ICSID Case No ARB/15/50, Decision on Respondent Request for Immediate Termination and Respondent Jurisdictional Objection based on Inapplicability of the Energy Charter Treaty to Intra-EU Disputes (7 May 2019) [231]–[232]; PL Holdings *Sàrl v Republic of Poland*, SCC Case No V2014/163, Judgment of the Svea Court of Appeal (22 February 2019) [175]–[176].

is introduced. This power, however, is only limited to the statements of claim and defence. In that case, the provision would not be of use in many instances in which the question of admitting illegal evidence arises long after the arbitration has commenced. In *Methanex v USA*, the illegal evidence was collected in the course of arbitral proceedings, not before them.[145] This was also the case in *Slovenian Border Dispute*, where the evidence obtained by tapping the arbitrator's phone was introduced long after the statements of claim and defence.[146]

Paragraph 3 of Article 24 of the UNCITRAL Rules parallels more the general power to order document production. Much like the ICSID Rules, the request can be given "at any time". The timeframe for producing such evidence, however, is different. A tribunal may order the evidence to be produced within a period of time that it will determine. Offering a longer period of time to the more vulnerable party would be a method of ensuring equality of arms. It also strikes at the very source of the problem in the admissibility of inappropriate evidence—one party may lack the resources or capacity to obtain similar evidence. Moving the deadlines may be one way of solving the problem.

For those reasons, the ICSID Rules might be better suited in preserving equality of arms by conferring a wider document production competence on tribunals.

An example of the application of document production by a tribunal under the ICSID Rules can be seen in the *Caratube v Kazakhstan* case.[147] Document production was used specifically to assist with the admissibility of illegal, publicly available evidence. At the request of the claimant, the tribunal would only order the production of documents which were not covered by client-attorney privilege. In doing so, the respondent was ordered to produce a list of documents which were covered by the privilege.[148] This gave the claimant an opportunity to comment on the admissibility, ensuring that the tribunal made a fairer decision. This is an example of the arbitrators balancing the interests between the parties. The case further explained the consequences of failing to produce requested evidence. The

[145] *Methanex* (n 23) pt II ch I [59].
[146] *Croatia* (n 30).
[147] *Caratube* (n 63).
[148] ibid [174].

tribunal stated that "negative inferences may be drawn as a result of a Party's failure to abide with their burden to produce specific, relevant documents".[149]

VI. Conclusion

The possibility to consider illegal and inappropriate evidence is by no means an invention of investment law. It appears that such practice originates from domestic laws, which are broad enough to allow its introduction, and likely also from public international law, despite the ICJ never addressing the issue directly.

The balancing exercise inevitably includes the arbitrators considering criminal issues, illegality, and impropriety. It seems accepted that such matters are arbitrable. Where particularly serious illegality is alleged, arbitrators should consider such arguments due to the existence of international public policy.

Investment tribunals approach the subject of admissibility carefully. Arbitrators will engage in a balancing or weighing exercise to decide whether the substantive relevance of the evidence outweighs procedural considerations originating from its illegality or the method of procurement. It is a case-by-case approach. For the *Methanex* tribunal, multiple acts of trespass over five months tilted the balance against admissibility. In the *EDF* arbitration, the doubtful authenticity of evidence led to its inadmissibility. The *Libonanco* tribunal suggested that evidence covered by the attorney-client privilege is not admissible regardless of materiality. In *Awdi*, evidence obtained from an ongoing domestic criminal case was inadmissible.

Although the introduction of tainted evidence is an uphill struggle that rarely succeeds, the opposite is true if the evidence is publicly available. Such documents would be generally admissible since there are fewer interests left to protect. In fact, the admissibility of public evidence is so evident that arbitrators have relied on WikiLeak documents on their own motion (*sua sponte*). The admissibility of public evidence, however, is not absolute. Their late introduction to the proceedings can prove fatal, which was the view taken by the *ConocoPhillips* tribunal, albeit with a strong dissent from one of the arbitrators. This approach is likely to be departed from in the future and limited to the facts of the case. Further, privileged attorney-client documents are also inadmissible, as was held in the *Caratube v Kazakhstan* arbitration. The approach of investment law towards admitting public evidence also diverges from public international law, where the ICJ generally ignores such evidence even despite parties consistently pleading them in submissions.

Turning to domestic law provides few answers. In English law, the admissibility of tainted evidence is possible, subject to public policy. Conversely,

[149] ibid [319].

the American 'fruit of the poisonous tree' renders evidence outright inadmissible. Neither of them nor any other approach can conclusively reflect a general principle of law. The American approach, however, has also been applied by the ECtHR and, arguably, by some investment tribunals. Further, no consistent principles can be derived from within and between the civilian and common law traditions.

The issue of the disadvantage which admitting tainted evidence creates against the opposite party remains. Tribunals attach importance to the weight of wrongfulness. The difficulty is that some types of wrongful conduct are not necessarily illegal in international law. Borderline cases exist—cases that would be clearly criminal under domestic law but not illegal under international law, including espionage. Further, the concept of unfriendly acts encompasses broader wrongfulness. Tribunals will also aggressively impede breaches of international public policy, such as corruption. This extent of illegality or impropriety will influence the weighing exercise of tribunals in admitting tainted evidence.

Further, investment law and arbitration rules developed principles of procedural fairness. These include the principle of equality of arms, good faith, and clean hands. These principles ensure some measure of having a level playing field between the parties. The non-observation of these principles may result in the unenforceability of arbitral awards for breach of procedural rules.

Tribunals also have other tools for the protection of procedural fairness other than refusing admission of evidence. Judicial assistance helps to evidence the criminality and illegality of conduct. They may also order the production of documents on their own motion. In other words, there are a variety of tools that may protect equality between parties on the one hand with the need for a just and full consideration of the evidence to resolve the dispute on the other.

This article analysed how tribunals, as well as select international courts, approach the issue of admissibility of tainted evidence. By distilling principles and distinguishing case law, the implications of the findings are practical. Admitting tainted evidence creates a domino effect, bringing into play other considerations. These include procedural fairness, good faith, and clean hands. Equally important are the questions of arbitrability of criminal laws and the ability of arbitrators to preserve the balance between parties.

The consideration of the latest developments in international law facilitates a doctrinal, normative discussion. Some authorities suggest that conduct which would be deemed domestically illegal would not be such if committed by a State. Others suggest that there is no doctrine of clean hands in international law. Such questions are entangled with the practical findings of this work and require deeper analysis. Future research is necessary to allow for a more harmonious development of investment law in the area. More importantly, there is a need for more case

law from international courts and tribunals. Until then, investment tribunals and practitioners will have to conduct a careful case-by-case balancing exercise of substantive and procedural fairness of tainted evidence.

Cambridge Law Review (2021) Vol VI, Issue i, 97–136

Reimagining a Centralised Cryptocurrency Regulation in the US: Looking through the Lens of Crypto-Derivatives

Sangita Gazi*

Abstract

Cryptocurrency as a reference asset in any derivative product ('crypto-derivatives') is opaque, complex, and unreliable. The pricing and settlement of crypto-derivatives have no standardized form and limit retail investors' ability to comprehend the terms of the product. Moreover, retail investors investing in crypto-derivatives are vulnerable to monetary losses due to cryptocurrency's highly speculative nature, price volatility, and spot market manipulation. Nonetheless, the regulatory approach to crypto-derivatives appears to vary from jurisdiction to jurisdiction. For instance, while regulators in the UK and the EU have recently banned crypto-derivatives to protect retail investors from the risk and volatility of the crypto-derivatives market, the US has taken a more hands-off approach. This paper presents a comparative analysis of the US regulatory responses to crypto-derivatives with specific references to the UK's and the EU's approaches and rationale towards crypto-derivatives regulations in their respective regions. Unlike the EU and UK, where the regulators introduced restrictive measures regarding cryptocurrency, the US regulatory efforts are primarily limited to interpreting cryptocurrency in light of the existing legal and regulatory framework. Further,

* Ph.D. Candidate, Faculty of Law, University of Hong Kong. Postgraduate Research Fellow, Asian Institute of Financial Law, University of Hong Kong. LL.M., Duke University School of Law; LL.M., University of Warwick. Former Assistant Legal Advisor at the US Department of Justice-OPDAT, US Embassy Dhaka, Bangladesh. I am grateful to Lee Reiners and Christopher Smith for their comments on earlier drafts. sangita.gazi@gmail.com

the regulatory approach in streamlining cryptocurrency and associated innovative products in the current framework inadequately encapsulates cryptocurrency's susceptibility to spot market manipulation and its potential to jeopardize investors' interests. Hence, it is paramount that the US enact comprehensive cryptocurrency regulation that recognizes the novelty of cryptocurrencies' market risks and introduces a robust regulatory infrastructure to limit market manipulation in the cryptocurrency spot market vis-à-vis the crypto-derivatives market. The paper envisions a cryptocurrency regulation that includes: (i) a centralised cryptocurrency trading platform; (ii) a mandatory registration requirement for all cryptocurrency exchanges and; (iii) a federal cryptocurrency agency. The paper suggests that with a degree of centralisation, a federal cryptocurrency agency is likely to establish the desired visibility into the cryptocurrency spot and an effective oversight mechanism that would eventually help curb market manipulation and restore investor confidence.

Keywords: crypto-derivatives, cryptocurrency, price volatility, investor protection, regulation.

I. Introduction

> "What I'm concerned about at the moment is if it can be reasonably demonstrated that the underlying trading is generally not manipulated, it's happening on reliable venues with good rules".[1]

On 6 October 2020, the UK Financial Conduct Authority ('FCA') prohibited the sale of cryptocurrency-related derivatives ('crypto-derivatives') to retail investors on the ground that cryptocurrency as a reference asset is an unreliable basis for valuation of these derivatives products.[2] The FCA concluded that crypto-derivatives, especially in the form of contract for difference ('CFD') and exchange-traded notes ('ETNs'), are ill-suited for retail consumers because of the harm they pose.[3] In Europe, the European Securities and Markets Authority ('ESMA') has also been looking to curb crypto-derivatives trading as these products are risky,

[1] Jay Clayton (SEC Chairman) quoted in Helen Partz,'SEC Chairman Highlights Investor Protection in Regard to Bitcoin ETF' (Cointelegraph, 14 March 2019) <https://cointelegraph.com/news/sec-chairman-highlights-investor-protection-in-regard-to-bitcoin-etf> accessed 23 February 2021.

[2] Financial Conduct Authority, 'Prohibiting the Sale to Retail Clients of Investment Products that Reference Cryptoassets' (2020) PS20/10 <www.fca.org.uk/publication/policy/ps20-10.pdf> accessed 24 February 2021.

[3] ibid.

speculative, and expose consumers to potentially huge losses.[4] Both regulators seem to have three main reasons for banning the sale of crypto-derivatives to retail investors: first, cryptocurrencies' extreme volatility as a reference asset; second, the prevalence of rampant market abuse, price manipulation, and security breaches in the cryptocurrency spot market and; investors' significant lack of understanding of these complex derivatives products.[5]

While regulators have initiated a broader crackdown in the UK and the EU to protect retail investors from the crypto-derivatives market's abuse and manipulation, the US regulators chose to go the opposite direction. In 2014, the Commodity Futures Trading Commission ('CFTC') approved TeraExchange, a bitcoin-derivatives exchange, to self-certify bitcoin swaps allowing investors to trade dollar-dominated bitcoin currency swaps.[6] In the following year, the CFTC classified bitcoin as a commodity in its order against Coinflip Incorporated, a bitcoin trading platform, and thus ensured its entrance into the traditional derivatives market just like other commodities.[7] Since then, several crypto-derivatives have proliferated in the market. The Chicago Mercantile Exchange ('CME') and the Chicago Board Options Exchange ('CBOE') first launched cash-settled bitcoin futures in December 2017.[8] The Intercontinental Exchange ('ICE') introduced physically-settled bitcoin futures and bitcoin options in September and October 2019, respectively.[9] Following the CFTC's announcement that the 'Ethereum'

4 European Securities and Markets Authority, 'ESMA Agrees to Prohibit Binary Options and Restrict CFDs to Protect Retail Investors' (27 March 2018) <www.esma.europa.eu/press-news/esma-news/esma-agrees-prohibit-binary-options-and-restrict-cfds-protect-retail-investors> accessed 24 February 2021.

5 Previously, South Korea banned bitcoin futures trading following its initial ban on Initial Coin Offerings (ICOs) in 2017; see David Dinkins, 'In Unexpected Move, South Korean Regulator Suddenly Bans Bitcoin Futures Trading' (Cointelegraph, December 2017) <https://cointelegraph.com/news/in-unexpected-move-south-korean-regulator-suddenly-bans-bitcoin-futures-trading> accessed 11 May 2020.

6 Michale Casey, 'TeraExchange Unveils First U.S. Regulated Bitcoin Swaps Exchange' (The Wall Street Journal, 12 September 2014) <www.wsj.com/articles/teraexchange-launches-bitcoin-derivatives-exchange-1410543989> accessed 11 November 2020.

7 CFTC v Coinflip Inc [2015] CFTC Docket No. 15-29. In this case, the CFTC held that bitcoin and other virtual currencies fall within Section1(A)(9) of the Commodity Exchange Act, as the definition of "commodity" shall include "all services, rights, and interests in which contracts for future delivery are presently or in the future dealt in". Therefore, any company offering bitcoin derivatives must comply with the CFTC laws, rules, and regulations.

8 Evelyn Cheng, 'Bitcoin Debuts on the World's Largest Futures Exchanges, and Prices Fall Slightly' (CNBC, 18 December 2019) <www.cnbc.com/2017/12/17/worlds-largest-futures-exchange-set-to-launch-bitcoin-futures-sunday-night.html> accessed 11 November 2020.

9 Ryan Brown, 'NYSE Owner ICE Launches Deliverable Bitcoin Futures Contracts' (CNBC, 23 September 2019) <www.cnbc.com/2019/09/23/nyse-owner-ice-launches-deliverable-bitcoin-futures-contracts.html> accessed 11 November 2020.

cryptocurrency is a commodity,[10] Eris Exchange ('ErisX') launched Ethereum-based physically settled futures contracts on 11 May 2020.[11]

The US crypto-derivatives and perpetual swap market cap stands at $319.11 billion.[12] As the market grows, the crypto-derivatives market's concerns are also emerging, as unregulated online exchanges, and brokerage firms offering cryptocurrency trading products are susceptible to spot market,[13] manipulation,[14] and cyber-attacks.[15] In support of the crypto-derivatives market, many argued that crypto-derivatives give institutional investors an efficient and confident way to hedge risk. However, this argument could be far from reality as two major global regulators — the FCA and the ESMA — view crypto-derivatives as harmful to retail investors due to its opaque and uncertain nature. From a regulatory standpoint, the CFTC's approach in regulating

[10] William Foxley, 'CFTC Chairman Confirms Ether Cryptocurrency is a Commodity' (Coindesk, 10 October 2019) <www.coindesk.com/cftc-chairman-confirms-ether-cryptocurrency-is-a-commodity> accessed 11 November 2020.

[11] Nikhilesh De, 'ErisX Announces Launch of First US Ether Futures Contracts' (Coindesk, 11 May 2020), <www.coindesk.com/erisx-announces-launch-of-first-us-ether-futures-contracts> accessed 11 November 2020.

[12] CoinMarket Cap, 'Cryptocurrency Derivatives and Perpetual Swap Markets' <https://coinmarketcap.com/derivatives/> accessed 20 February 2021. It is important to mention that on 17 February 2021, the market cap was $134.21 billion, which means the market grew more than twice in 72 hours.

[13] Spot market refers to the trading and prices of cryptocurrency in any exchange. Unlike traditional spot markets for commodities, cryptocurrencies have its own niche spot markets. There are over 300 cryptocurrency exchanges, and some Over the Counter markets, that constitute the cryptocurrency spot markets.

[14] A few notable examples of regulators' enforcement action against market manipulation in the cryptocurrency spot market are: (1) On 20 October 2020, the CFTC announced that the US District Court for the Southern District of New York ordered a person to pay $7.4 million for committing a multi-million-dollar bitcoin fraud (CFTC Release No. 8272-20); (2) On 18 June 18 2019, the CFTC charged Control-Finance Limited, a purported bitcoin trading and investment company, and its principal, Benjamin Reynolds, for fraudulently obtaining and misappropriating $147 million worth of bitcoins from more than 1,000 customers (CFTC Release No. 7938-19); (3) On 23 July 2018, the Federal Court order a commodity pool operator and its principal to pay more than $1.9 million in connection with a bitcoin and binary options fraud scheme (CFTC Release No. 7760-18); and (4) In another case, a New York Federal Court ordered a trading firm and its CEO to pay more than $2.5 million for operating a bitcoin ponzi scheme (CFTC Release No. 7831-18). See also, Neil Gandal, JT Hamrick, Tyler Moore, and Tali Oberman, 'Price Manipulation in the Bitcoin Ecosystem' [2018] 95 Journal of Monetary Economics 86.

[15] In a study about market manipulation behaviour in the cryptocurrency exchanges, the evidence showed that the biggest cryptocurrency exchange, Mt. Gox, was engaged in bitcoin price manipulation before it was hacked in 2013, that led the exchange to file for bankruptcy in 2014. A subsequent data leak revealed that a significant number of trades took place at rates that were far higher or far lower than the reference price. The findings also demonstrated that these abnormal transactions took place between two accounts (presumably belonging to Mt. Gox itself), which artificially inflated the daily bitcoin trade volume to manipulate the price. See Weili Chen and others, 'Market Manipulation of Bitcoin: Evidence from Mining the Mt. Gox Transaction Network' <https://arxiv.org/abs/1902.01941> accessed 23 February 2021.

crypto-derivatives through self-certification (like traditional derivatives products) inadequately encapsulates the cryptocurrency spot market's inherent risks of opacity, price volatility, and exposure to market manipulation.[16] Such inadequacy is embedded in the CFTC's two contrasting positions. In a traditional commodity derivatives market, the CFTC has the power and capacity to both oversee the commodity spot markets and, therefore, to take enforcement actions against any abusive and manipulative behaviour that is detrimental to the investors' interests.[17] However, concerning crypto-derivatives, the CFTC's oversight mechanism over the cryptocurrency spot market is debatable as the CFTC has, on repeated occasions, appeared to have conflicting opinions regarding such power. Furthermore, market participants in the spot market operate with the assumption that the spot market is beyond the CFTC's regulatory perimeter. Hence, in the absence of any regulatory clarity and with the CFTC's questionable oversight mechanism, the spot market could be a means to incentivise a bad actor to jeopardise crypto-derivatives markets' integrity and thereby undermine the retail investors' confidence.

In addition, investor protection is a grey area in the US cryptocurrency regulatory regime. As an example, although an Initial Coin Offering ('ICO') may be a security,[18] it is uncertain whether all investments are protected under the Security Investment Protection Act ('SIPA') given that the SEC has also determined that not all digital tokens are securities as depending on the degree of decentralization of platform the offering takes place, a coin or token may fall outside the definition of a security.[19] Similarly, investor protection in the crypto-derivatives market also remains vague as the effectiveness of the CFTC's regulatory and oversight mechanisms in preventing manipulation in the cryptocurrency spot market is questionable. Most investors, especially retail investors, lack an understanding of the complexity of cryptocurrency pricing and thus tend to treat cryptocurrency

[16] For a detailed discussion regarding the impact of self-certification on bitcoin futures, see Lee Reiners, 'Bitcoin Futures: Self-certification to System Risk' [2019] 23 North Carolina Bank Institute 61.

[17] Enacted after the financial crisis of 2007, the Dodd-Frank Act authorises the CFTC to bring the OTC under a broader regulatory purview, and thereby establish a direct visibility into the commodity spot market.

[18] Securities and Exchange Commission, 'Report of Investigation Pursuant to Section 21(a) of the Securities Exchange Act of 1934: The DAO' (2017) Release No. 81207 < www.sec.gov/litigation/investreport/34-81207.pdf> accessed 11 March 2021.

[19] ibid. The SEC is of the view that "[w]hether or not a particular transaction involves the offer and sale of a security—regardless of the terminology used—will depend on the facts and circumstance, including the economic realities of the transaction" (n 18) 17–18. See also, William Hinman, 'Digital Asset Transactions: when Howey Met Gary (Plastic)' (US Securities and Exchange Commission, 14 June 2018) <www.sec.gov/news/speech/speech-hinman-061418> accessed 23 November 2020.

trading like gambling.[20] Furthermore, the complex pricing combined with extreme price volatility gives main-street investors an incentive to speculate the cryptocurrency price.[21] Finally, the cryptocurrency market size incentivises the institutional money to flow into the new cryptocurrency-based economy, and therefore, calls for regulators' vigilance.[22]

Against this background, this paper puts forth a comparative analysis of the US regulatory responses to crypto-derivatives with specific references to the UK's and the EU's approaches and motives towards crypto-derivatives regulations in their respective regions. It discusses that the UK and the EU regulators primarily focus on protecting retail investors from monetary losses arising from investment in crypto-derivatives products. In contrast, the US regulatory efforts are limited to interpreting cryptocurrency in light of the existing legal and regulatory framework. In the absence of CFTC's oversight over the cryptocurrency spot market, regulating cryptocurrency-related products under the traditional laws undermines cryptocurrency's novel risks of price volatility and susceptibility to spot market manipulation. By comparing the US crypto-derivatives regulation with that of the UK's and the EU's, this paper does not necessarily suggest that the US should follow either of these jurisdictions and issue an outright ban — permanent or temporary — on crypto-derivatives. What the paper emphasises on is that the US incorporates a robust crypto-derivatives regulation that captures this novel product's complex risks and uncertainties, and tailors the regulation to protect the 'Main Street' investors' interest. It nevertheless explores the possibility of imposing an outright ban on crypto-derivatives like the UK and concludes that such a ban on the crypto-derivatives market is likely to jeopardise financial innovation growth

[20] In a market survey conducted in the UK, the majority's perception regarding cryptocurrency is, it is akin to betting. A study showed that cryptocurrency trading is linked with problematic gambling. See Yessi Bello Perez, 'Problem Gamblers More Likely to Obsessively Trade Cryptocurrency, Research Finds' (The Next Web.com, 11 March 2019) <https://thenextweb.com/cryptocurrency/2019/03/11/problem-gamblers-more-likely-to-obsessively-trade-cryptocurrency-research-finds/> accessed 11 November 2020.

[21] The cryptocurrency hedge-funds are betting on bitcoin's price. See Vincent Mislos, 'Bitcoin Price will Hit $20,000 This Year because of "Liquidity Pump", Says Novogratz' (International Business Times, 30 July 2020) <www.ibtimes.com/bitcoin-price-will-hit-20000-year-because-liquidity-pump-says-novogratz-3019421> accessed 23 November 2020.

[22] In the age of digital communications system and cryptocurrency-based financial system, the states have to reimagine their roles in protecting financial stability and hence, redesign the financial regulatory structure. For an academic discussion on regulation in the context of an emerging lex cryptographica financiera, see Jason Grant Allen and Rosa María Lastra, 'Border Problems: Mapping the Third Border' [2020] 83 Modern Law Review 505.

in the US.[23] Furthermore, some cryptocurrency trading platforms are already complying with the existing laws and regulations, and an outright ban will set them back. Therefore, to protect market integrity and safeguard retail investors' interest, it is paramount that the cryptocurrency spot market be regulated.

The proposition this paper puts forward is that without an effective and robust crypto-regulation with a certain degree of centralization, the market manipulation in the spot markets will continue. This eventually hurts the crypto-derivatives market, and hence, requires a parallel discussion. Furthermore, amidst the COVID-19 pandemic, digital finance, including cryptocurrency, has been witnessing accelerated growth.[24] Big corporations, such as Tesla, BNY Mellon and Mastercard are reported to have invested in cryptocurrency.[25] Facebook is testing the launch of Diem (formerly known as 'Libra') — a global stablecoin[26] that is designed to be pegged to US dollar.[27] Such expansion of digital finance and the use of cryptocurrencies among big corporations compelled regulators worldwide

[23] In November 2019, the CFTC Chairman, Health P. Tarbert, expressed his intention to make the US a leading nation in the field of blockchain and digital assets. So, it is highly unlikely that the US regulators will take any decision of putting an outright ban on the crypto-derivatives. See Miranda Wood, 'CFTC Chairman Wants to Lead in Blockchain' (Ledger Insight, 21 November 2019) < https://www.ledgerinsights.com/cftc-chairman-us-blockchain/> accessed 11 May 2021.

[24] For reference, see The World Bank, 'Fintech Market Reports Rapid Growth During COVID-19 Pandemic' (The World Bank, 3 December 2020) <www.worldbank.org/en/news/press-release/2020/12/03/fintech-market-reports-rapid-growth-during-covid-19-pandemic> accessed 21 February 2021. See also Chris Versace, Lenore Elle Hawkins and Mark Absy, 'The Rising Tide of Digital Currencies' (NASDAQ, 19 February 2021) <www.nasdaq.com/articles/the-rising-tide-of-digital-currencies-2021-02-19> accessed 21 February 2021.

[25] BBC, 'Elon Musk's Tesla Buys $1.5bn of Bitcoin Causing Currency to Spike' (BBC, 8 February 2021) <www.bbc.com/news/business-55939972> accessed 21 February 2021. See also, Penny Crossman '"Digital Assets are Here to Stay": BNY Mellon Embraces Crypto' (American Banker, 23 February 2021) <www.americanbanker.com/news/digital-assets-are-here-to-stay-bny-mellon-embraces-crypto> accessed 24 February 2021; Jenna Delport, 'Mastercard to Support Cryptocurrency Transactions on its Network' (The CNBC, 11 February 2021) <www.itnewsafrica.com/2021/02/mastercard-to-support-cryptocurrency-transactions-on-its-network/> accessed 24 February 2021.

[26] Stable coins are digital currencies pegged to fiat currencies or non-volatile assets or to fixed amounts of traditional monetary instruments. Stable coins came into cryptocurrency markets to resolve the problem the problem of cryptocurrencies' market volatility. For reference, see Aleksander Berensten and Fabian Schär, 'Stablecoins: The Quest for a Low-Volatility Cryptocurrency' in Antonio Fatas (ed.), The Economics of Fintech and Digital Currencies (CEPR Press 2019) 65–74.

[27] For an academic discussion on Libra and its impact on payment and monetary system landscape, see Dirk A. Zetzsche, Ross P. Buckley and Douglas W. Arner, 'Regulating LIBRA: The Transformative Potential of Facebook's Cryptocurrency and Possible Regulatory Responses' (forthcoming) Oxford Journal of Legal Studies <https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3414401> accessed 21 February 2021.

to reimagine the legislative actions required in addressing issues concerning cryptocurrency and stable coins.

The UK and the EU have moved towards implementing a coherent, robust, and uniform regulatory structure for the cryptocurrency industry and market participants that provides stringent protection measures to retail investors and consumers while supporting financial innovation and stability. The paper proposes that the US Congress enact a centralised, comprehensive cryptocurrency regulation ('crypto-regulation'), recognising the novelty of the cryptocurrencies' market risks, and introducing effective regulatory treatments, to curb market manipulation in the cryptocurrency spot market *vis-à-vis* crypto-derivatives market. It is paramount that such crypto-regulation is not fragmented,[28] but rather centralised, conferring specific jurisdiction relating to cryptocurrency and cryptocurrency-related financial products on a single US regulatory body. To this end, this paper envisions a centralised US crypto-regulation that would include: (i) centralization of cryptocurrency trading platforms; (ii) a mandatory registration requirement for all cryptocurrency exchanges and; (iii) a single federal cryptocurrency agency having exclusive jurisdiction over cryptocurrencies and oversight authority on the cryptocurrency spot market.

II. Regulatory Framework of Crypto-Derivatives in the EU and the UK

Unlike the US, regulators worldwide are sceptical about crypto-derivatives — mainly because of their extremely volatile, speculative, and high-leverage nature.[29] The complexity of crypto-derivative' products and investors' lack of understanding regarding the risks associated with come with a high likelihood of losing money.[30] Therefore, among other major regulators, the EU and the UK have

[28] In the US, the major regulators concerning financial products are the SEC (for securities and security-based derivatives) and the CFTC (for commodity and financial derivatives). Under the existing legal framework, the SEC regulates the digital tokens and ICOs which they determine as securities, whereas the CFTC regulates derivatives products where cryptocurrency is used as a reference asset. The proposed crypto-regulation will obliterate this division between the SEC's and the CFTC's mandate over cryptocurrency and establish a centralised cryptocurrency regulatory body.

[29] For instance, in May 2019, Japan asked bitFlyer to reduce leverage for its perpetual swap product. See Emmanuel Goh, 'Crypto Derivatives: A Corner of the Market or the Market Itself?' (Coindesk, 16 November 2019) <www.coindesk.com/crypto-derivatives-a-corner-of-the-market-or-the-market-itself> accessed 11 November 2020.

[30] ibid.

evaluated their regulations of crypto-derivatives on the ground that investment in these products hurt retail investors.

A. The EU

In the EU, the derivatives markets are regulated by two central EU regulations, namely, the European Market Infrastructure Regulation ('EMIR')[31] and the Markets in Financial Instruments Directive ('MiFID II'),[32] alongside the Markets in Financial Instruments Regulation ('MiFIR').[33] Under these regulations, the ESMA is the independent authority for market supervisory and law enforcement of the EU derivatives markets.[34] The ESMA is authorised to clear all eligible derivatives contracts and is responsible for trade repositors' surveillance across the EU.[35] Besides derivatives market, the ESMA is also responsible for promoting

[31] Regulation (EU) No 648/2012 on OTC derivatives, central counterparties and trade repositories [2012] OJ L201/1. In the wake of the US financial crisis 2007, the EU OTC derivatives markets also went through transformation. As a result, in 2012, the European Commission promulgated the EMIR to include and regulate broad range of OTC derivatives across various asset classes, central counterparties and trade repositories.

[32] Directive 2014/65/EU on markets in financial instruments [2014] OJ L173/349 ("MiFID II"). Previously, in 2004, the European Commission adopted the Markets in Financial Instruments Directive ("MiFID") which was in force between 2007 and 2018. In 2014, the Commission revised the MiFID framework and adopted new rules composed of a directive—MiFID II and a regulation, MiFIR. Under the MiFID, the investors are categorised in three separate groups: professional clients, retail clients, and eligible counterparties. The aim of such division among investors is to reflect the necessity of different level of protection an investor may need. According to the classification, the retail investors needed the highest level of protection and comprehensive information that are required for them to understand the risks associated with a specific investment product and transaction. For reference on MiFID II's impact on investor protection, see Christos Gortsos, 'Stricto Sensu Investor Protection under MiFID II: A Systemic Overview of Articles 24–30' (1st ed., Cambridge Scholars Publishing 2018).

[33] Regulation (EU) No 600/2014 on markets in financial instruments [2014] OJ L173/84.

[34] Philipp Maume and Mathias Fromberger, 'Regulation of Initial Coin Offerings: Reconciling US and EU Securities Laws' [2019] 19 Chicago Journal of International Law 548.

[35] European Commission, 'Derivatives / EMIR'(European Commission) <https://ec.europa.eu/info/business-economy-euro/banking-and-finance/financial-markets/post-trade-services/derivatives-emir_en> accessed 23 February 2021.

"supervisory convergence and the consistent application of market rules"[36] within the EU.

The ESMA first stepped into the cryptocurrency world by expressing its view on cryptocurrency token offering, by way of ICOs, in November 2017.[37] Although the ESMA's statement regarding token sales was vague as it largely leaves the burden on the firms and investors for their activities,[38] the ESMA has quite a strong position in regulating the EU's crypto-derivatives market. In a Call for Evidence Report[39] issued in January 2018,[40] the ESMA announced that crypto-derivatives, which are in the form of CFDs[41] and BOs,[42] should be subject

[36] ibid.

[37] European Securities and Markets Authority, 'The ESMA alerts firms involved in Initial Coin Offerings (ICOs) to the need to meet relevant regulatory Requirements' (13 November 2017) ESMA50-157-828 ("first statement") <http://perma.cc/A4BP-9QS4> accessed 11 November 2020. This statement was published alongside with the ESMA's statement made towards the investors warning the risks involved with the ICOs. See European Securities and Markets Authority, 'ESMA alerts investors to the high risks of Initial Coin Offerings (ICOs)' (13 November 2017) ESMA50-157-829 <www.esma.europa.eu/sites/default/files/library/esma50-157-829_ico_statement_investors.pdf> ("second statement") accessed 20 February 2021.

[38] Ibid. In the first statement, the ESMA delivers a blanket statement stating, "Firms involved in ICOs must give careful consideration as to whether their activities constitute regulated activities", without elaborating or giving precise guidelines as to what may be construed as "regulated activities". The second statement further states that, "[d]epending on how they are structured, ICOs may fall outside of the scope of the existing rules and hence outside of the regulated space. However, where the coins or tokens qualify as financial instruments it is likely that the firms involved in ICOs conducted regulated investment activities" and hence, should be subject to the EU securities laws and regulations. Without delineating specific conditions or requirements, the ESMA appears to leave the burden of compliance on the firms offering digital tokens and ICOs.

[39] European Securities and Markets Authority, 'Call for Evidence: Potential Intervention Measures on Contracts for Differences and Binary Options to Retain Clients' (18 January 2018) ESMA35-43-904 <www.esma.europa.eu/sites/default/files/library/esma35-43-904_call_for_evidence_-_potential_product_intervention_measures_on_cfds_and_bos_to_retail_clients.pdf> accessed 11 November 2020.

[40] European Securities and Markets Authority, 'ESMA Consults on Potential CFDs and Binary Options Measures to Protect Retail Investors' (European Securities and Markets Authority, 18 January 2018) <www.esma.europa.eu/press-news/esma-news/esma-consults-potential-cfd-and-binary-options-measures-protect-retail> accessed 11 November 2019.

[41] A CFD is defined as "a derivative other than an option, future, swap, or forward rate agreement, the purpose of which is to give the holder a long or short exposure to fluctuations in the price, level or value of an underlying, irrespective of whether it is traded on a trading venue, and that must be settled in cash at the option of one of the parties other than by reason of default or other terminational event". See (n 40) 4.

[42] A BO is defined as "a derivative that meets the following conditions: (a) it must be settled in cash or may be settled in cash at the option of one of the parties other than by reason of default or other terminational event; (b) it only provides for payment at its close-out or expiry; and (c) its payment is limited to: (i) a predetermined fixed amount if the underlying of the derivative meets one or more predetermined conditions; and (ii) zero or another predetermined fixed amount if the underlying of the derivative does not meet one of more predetermined conditions". See European Securities and Markets Authority (n 40) 4.

to strict legal scrutiny alleging that these derivatives products are speculative and volatile, exposing investors to potentially significant monetary loss.[43] The ESMA further called for responses from market participants regarding possible measures to regulate crypto-derivatives.[44] After considering all responses and concerns, the ESMA, according to Art. 40 of MiFIR,[45] adopted restrictive product invention measures in relation to CFDs and BOs.[46] The intervention measures include: (1) a prohibition on the marketing, distribution, or sale of BOs and (2) a restriction on the marketing, distribution, or sale of CFDs to retail investors.[47] In adopting these restrictive measures, the ESMA noted that:

> "[…] CFDs are complex products. The pricing, trading terms, and settlement of such products is not standardized, impairing retail investors' ability to understand the terms of product. In addition, CFD providers often require investors to acknowledge that the reference prices used to determine the value of a CFD may differ from the price available in the respective market where the underlying is traded, making it difficult for retail investors to check the accuracy of the prices received from the CFD provider".[48]

It also noted that cryptocurrency is an immature asset class that poses "separate and significant concerns".[49] Therefore, retail investors hardly understand the risk of speculation on crypto-derivatives products. The ESMA from time to

[43] European Securities and Markets Authority, 'Additional Information on the Agreed Product Intervention Measure Relating to Contract for Differences and Binary Options' (European Securities and Markets Authority, 27 March 2018) <www.esma.europa.eu/sites/default/files/library/esma35-43-1000_additional_information_on_the_agreed_product_intervention_measures_relating_to_contracts_for_differences_and_binary_options.pdf> accessed 11 November 2020.

[44] ibid.

[45] Article 40 of MiFIR (n 33). It permits the ESMA to temporarily prohibit, restrict marketing, distribution, or sale of certain financial instruments on grounds of investor protection and market integrity.

[46] European Securities and Markets Authority (n 4).

[47] ibid. The restrictions on BOs came in effect from 1 July 2018, whereas the restrictions on CFDs came in effect from 1 August 2018 (the restrictions on CFDs are renewable).

[48] (n 44).

[49] ibid 5. The ESMA also states: "[…] CFDs with cryptocurrencies as an underlying raise separate and significant concerns as CFDs on other underlyings. Cryptocurrencies are a relatively immature asset class that pose major risks for investors. ESMA and NCAs have significant concerns about the integrity of the price formation process in underlying cryptocurrency markets, which makes it inherently difficult for retail clients to value these products…".

time extended its restriction on CFDs and BOs.[50] In renewal notices, the ESMA reiterated its concern over investor protection related to the sale of CFDs and BOs to retail clients.[51]

B. The UK

The FCA, that regulates the UK financial services industry, has imposed strict regulatory measures in the UK crypto-derivatives market.[52] Before finalizing the outright ban on crypto-derivatives,[53] the FCA first proposed a temporary ban on crypto-derivatives and ETNs in 2019, on the ground that crypto-derivatives products were ill-suited to retail customers who are unable to assess the value and risks of derivatives or ETNs reliably.[54]

To assess the trend of investors' increasing interest and its correlation with cryptocurrencies' price instability, the FCA evaluated the price of bitcoin and Ethereum, and Google trends data between 2018 and 2019. By doing so, it was demonstrated that retail investors' interests are strongly "correlated to the increasing price and trading volumes of bitcoin"[55] as well as ethereum. The FCA

[50] See European Securities and Markets Authority, 'ESMA to Renew Restriction on CFD for a Further Three Months' (European Securities and Markets Authority, 28 September 2018) <www.esma.europa.eu/sites/default/files/library/esma71-99-1041_-_esma_to_renew_restriction_on_cfds_for_a_further_three_months.pdf> accessed 11 November 2020. European Securities and Markets Authority, 'ESMA to Renew Restrictions on CFDs for a Further Three Months from 1 May 2019' (European Securities and Markets Authority, 27 March 2019) < https://www.esma.europa.eu/press-news/esma-news/esma-renew-restrictions-cfds-further-three-months-1-may-2019> accessed 11 May 2021. See also European Securities and Market Authority, 'ESMA Renews Binary Options Prohibition for a Further Three Months from 2 January 2019' (European Securities and Markets Authority, 09 November 2018) < https://www.esma.europa.eu/sites/default/files/library/esma71-99-1057_-_esma_renews_binary_options_prohibition_for_a_further_three_months_from_2_january_2019.pdf> accessed 11 May 2021. See also European Securities and Markets Authority, 'ESMA Renews Binary Options Prohibition for a Further Three Months from 2 April 2019' (European Securities and Markets Authority, 18 February 2019) < https://www.esma.europa.eu/press-news/esma-news/esma-renews-binary-options-prohibition-further-three-months-2-april-2019> accessed 11 May 2021.

[51] ibid.

[52] In April 2018, the FCA released additional guidance regarding derivative contracts on cryptocurrencies, making it clear that derivatives on crypto tokens are transferable securities and that providing financial services in this regard require formal authorization. See Financial Conduct Authority, 'FCA proposes ban on sale of crypto-derivatives to retail consumers' (Financial Conduct Authority, 3 July 2019) <www.fca.org.uk/news/press-releases/fca-proposes-ban-sale-crypto-derivatives-retail-consumers> accessed 11 November 2020.

[53] Financial Conduct Authority (n 2).

[54] Financial Conduct Authority, 'Prohibiting the Sale to Retail Client of Investment Products that Reference Cryptoassets' (July 2019) CP19/22 <www.fca.org.uk/publication/consultation/cp19-22.pdf> accessed 24 February 2021.

[55] Financial Conduct Authority (n 2) 9, 11.

is of the view that the data further demonstrated investors' speculative behaviour over a price-boom in cryptocurrency, rather than their ability to reliably and consistently assess the intrinsic value of cryptocurrency, or the derivatives that use cryptocurrency as a reference asset.[56]

Therefore, in framing the grounds for banning crypto-derivatives, the FCA's central focus was protecting retail investors from monetary losses.[57] The regulators were concerned that retail investors could be hurt because of: (i) the opacity and complexity of cryptocurrency as reference assets;[58] (ii) retail investors' lack of understanding and consequent inability to make an informed investment decision on crypto-derivatives[59] and; (iii) the cryptocurrency as a reference asset is highly speculative,[60] volatile,[61] and susceptible to sudden price drops and abrupt price dislocation.[62] Allowing crypto-derivatives to grow in the retail market might create a perception among retail investors that these products are suitable and appropriate investment products. In considering the proportionality of a ban on crypto-derivatives, the FCA invoked Art. 42 of MiFIR[63] and Art. 21(2) of the Delegated Regulation of MiFIR,[64] and determined that a permanent ban on crypto-derivatives is an appropriate measure to secure the interest of retail investors. During the interim phase, the FCA considered other less interventionist

[56] ibid 8.

[57] Financial Conduct Authority (n 54).

[58] ibid 14.

[59] ibid.

[60] Shay-Kee Tan, Jennifer So-Kuen Chan and Kok-Haur Ng, 'On the Speculative Nature of Cryptocurrencies: A Study of Garman and Klass Volatility Measure' [2020] 32 Finance Research Letters.

[61] Bitcoin is 26 times as volatile than S&P 500. See C. Baek and M. Elbeck, 'Bitcoin as an investment or speculative: A first look' 22 [2015] 1 Applied Economies Letter 34.

[62] Financial Conduct Authority (n 2) 9, citing CP19/22 (n 57).

[63] Article 42 of the MiFIR (n 33) provides a competent authority with the power to prohibit or restrict: "(a) the marketing, distribution or sale of certain financial instruments or structured deposits or financial instruments or structured deposits with certain specified features; or (b) a type of financial activity or practice."

[64] Commission Delegated Regulation (EU) 2017/567 supplementing Regulation (EU) No 600/2014 of the European Parliament and of the Council with regard to definitions, transparency, portfolio compression and supervisory measures on product intervention and positions [2016] OJ L87/90. Article 21(2) lays down the factors and criteria to be assessed by a competent authority to "determine the existence of a significant investor protection concern or a threat to the orderly functioning and integrity of financial markets or commodity markets […]".

approaches such as "do nothing"[65] or "provide further consumer warnings".[66] Nevertheless, it concluded that "any remedy other than a ban on the sale to retail clients would fall short of adequately reducing the harms to consumers and risks identified".[67]

The FCA's efforts to regulate crypto-derivatives were not unopposed.[68] The FCA's position was challenged on the ground that "an outright ban would affect its members who are already in compliance with a slew of regulatory standards".[69] However, the FCA continues to maintain its position on the matter to protect retail investors, stating, "a ban on crypto-derivatives could lead to a $96 million haircut in harm done to retail traders per year".[70]

III. Regulatory Framework of Crypto-Derivatives in the US

In the US, the Commodity Exchange Act ('CEA') and the Commodity Futures Trading Commission Rules ('CFTC Rules') regulates the trading of derivatives contracts (including futures, options, and swaps), and the CFTC supervises the commodity and derivatives markets. The CEA, that is the primary statute governing the laws and regulations of the US derivatives market, defines "commodity" to include agricultural products, "all other goods and articles", and "all services, rights, and interests", in which "contracts for future delivery are presently or in the future dealt in".[71] In 2015, the CFTC assumed that certain virtual currencies, such as bitcoin and litecoin, are commodities, and should be regulated

65 Financial Conduct Authority (n 54) 24. The FCA is of the opinion that a "do nothing" approach does not address the fundamental product flaws or address the significant harm to consumers posed by these products. Existing disclosure obligations and appropriateness tests are unlikely to be effective in conveying the risks to retail clients. Continuing to allow the offer of these products by firms with FCA authorization may also give retail investors a false sense of security by contrast to the underregulated nature of the underlying.

66 Financial Conduct Authority, 'Consumer Warning About the Risks of Investing in Cryptocurrency CFDs' (Financial Conduct Authority, 3 July 2019) <www.fca.org.uk/news/news-stories/consumer-warning-about-risks-investing-cryptocurrency-cfds> accessed 11 November 2020.

67 Financial Conduct Authority (n 54) 24.

68 Steve Kaaru, 'CoinShares Wants Users to Take Action against UK Crypto Assets Ban' (Coingeek, 29 September 2019) <https://coingeek.com/coinshares-wants-users-to-take-action-against-uk-crypto-assets-ban/> accessed 11 November 2020.

69 Osato Avan Nomayo, 'Crypto Derivatives Ban: The UK Govt Won't Interfere with FCA' (Blockonomi, 29 October 2019), <https://blockonomi.com/crypto-derivatives-ban-uk-govt-wont-interfere-with-fca/> accessed 11 November 2020.

70 ibid.

71 Section 1a (9), Commodity Exchange Act.

by the CFTC.[72] Besides, multiple federal courts also held that virtual currencies are commodities as per the CEA.[73] Hence, crypto-derivatives — such as bitcoin-futures, swaps, and options — fall within the CFTC's regulatory perimeter.[74]

In December 2017, the CFTC permitted futures exchanges to apply the self-certification process for bitcoin-futures and binary options under §7(a)(2) of the CEA.[75] However, despite the CFTC's attempt to normalise crypto-derivatives in the existing legal and regulatory framework, these derivatives products pose numerous risks to retail consumers. Its lack of direct oversight on the cryptocurrency spot market poses a significant challenge to regulate market manipulation, that has adverse impacts on crypto-derivatives investors. In July 2018, Daniel Grofine, then Director of the CFTC's fintech initiative ('LabCFTC'), shared similar concerns on the issue of cryptocurrencies and digital assets during his testimony before the US House Committee on Agriculture.[76] He warned that while many things could be commodities, the CFTC's direct oversight on the commodity spot market is essential to bring those commodity-built futures, swaps, and options within its regulatory perimeter.[77] The current regulatory approach should focus on bringing clarity and certainty to the market, and any "hasty regulatory pronouncements are likely to […] have unintended consequences, or fail to capture important nuance regarding the structure of new products and models".[78]

Currently, there are several crypto-derivatives products available to US retail customers.[79] The ICE launched its first bitcoin-settled futures, the Bakkt futures,

[72] CFTC v Coinflip Inc (n 7). See also, Matt Clinch, 'Bitcoin Officially Becomes a Commodity' (CNBC, 15 September 2018) <www.cnbc.com/2015/09/18/bitcoin-now-classed-as-a-commodity-in-the-us.html> accessed 11 November 2020.

[73] CFTC v McDonnell [2018] 287 F Supp 3d 213 (EDNY 2018). See also CFTC v. My Big Coin Pay, Inc. [2018] 334 F Supp 3d 492 (D Mass 2018).

[74] See Houman B. Shadab, 'Regulating Bitcoin and Blockchain Derivative' (2020) NYLS Legal Studies Research Paper <www.cftc.gov/sites/default/files/idc/groups/public/@aboutcftc/documents/file/gmac_100914_bitcoin.pdf> accessed 11 November 2020 (discussing whether bitcoins fall within the definition of "commodity" under the Commodity Exchange Act (CEA), and therefore, derivatives contracts like futures, swaps, options that reference bitcoins are subject to regulation by the CFTC).

[75] Commodity Futures Trading Commission, 'A CFTC Primer on Virtual Currencies' (Commodity Futures Trading Commission, 17 October 2017) <www.cftc.gov/sites/default/files/idc/groups/public/documents/file/labcftc_primercurrencies100417.pdf> accessed 12 November 2020.

[76] Commodity Futures Trading Commission, 'Written Testimony of Daniel S. Gorfine before the US House Committee on Agriculture (Commodity Futures Trading Commission, 18 July 2018) <www.cftc.gov/PressRoom/SpeechesTestimony/opagorfine1> accessed 24 February 2021.

[77] ibid.

[78] ibid.

[79] For reference, see CoinMarketCap <https://cryptoderivatives.market/> accessed 24 February 2021.

in September 2019.[80] Three months later, the ICE introduced its monthly settled bitcoin options.[81] In January 2020, the CME started trading options on bitcoin futures.[82] ErisX launched ether-based physically settled futures contracts in May 2020.[83] In addition to bitcoin and ether derivatives products, the cryptocurrency industry will soon attempt to issue other cryptocurrency-based derivatives products. However, there are at least three regulatory issues with the CFTC's approach to approving these new crypto-derivatives.[84] First, the CFTC's traditional approach to regulating crypto-derivatives, primarily through the 'self-certification' process, is risky as the existing legal framework of 'self-certification' is not adequate to prevent price manipulation in the cryptocurrency spot market. Second, the CFTC's view on market manipulation in the cryptocurrency spot market contradicts the SEC's view on the same issue. Third, the CFTC is surprisingly numb to the suggestion that crypto-derivatives could jeopardise retail investors' interest, and such an approach deviates from the two major global regulators, that is the ESMA and the FCA.

A. The CFTC's self-certification process and the heightened review for crypto-derivatives undermine cryptocurrencies' market risks

The 'self-certification' process for derivatives contracts was introduced in 2000 by enacting the Commodity Futures Modernization Act.[85] Under this law, the CFTC permits the listing of a new futures contract if: (1) the exchange submits a written self-certification to the CFTC certifying that the contract complies with the CEA and CFTC regulations or (2) the exchange has voluntarily submitted the contract for CFTC approval.[86] Therefore, in the self-certification process, the exchanges themselves can verify that a new contract complies with the CEA's or the CFTC's requirements.[87] The designated contract markets ('DCMs') may

80 Brown (n 9).

81 Adam White, 'Expanding the Bakkt Bitcoin Product Complex: Bitcoin Options and Cash Settled Futures Now Available' (Bakkt Blog, 9 December 2019) <https://medium.com/bakkt-blog/expanding-the-bakkt-bitcoin-product-complex-68000faea6b3> accessed 11 November 2020.

82 CME Group <www.cmegroup.com/trading/bitcoin-futures.html> accessed 11 November 2020.

83 Brown (n 9).

84 This paper will not discuss systemic risk aspect of crypto-derivatives. This paper, however, admits that the crypto-derivatives connect regulated sectors, i.e., firms and financial institutions, with the unregulated underlying cryptocurrency markets. Therefore, any contagion created in the unregulated asset class may have a spill over impact on the regulated sector, that can give rise to systemic risk. For a discussion of crypto-derivatives' systemic risks, see Reiners (n 16).

85 Section 7(a)(2), Commodity Exchange Act.

86 Commodity Futures Trading Commission Regulation 40.2 (17 Code of Federal Regulations 40.2).

87 Reiners (n 16) 71.

also voluntarily submit new contracts for approval to the Commission and list the futures contract within twenty-four hours upon the CFTC's approval of the contract.[88] The CFTC's self-certification process has, however, been questionable since its introduction.[89] Between 2000 and 2017, data[90] suggests that the self-certification process facilitated the approval and listing of many complex exchange-traded commodity derivatives.[91] Such approval process often included an absence of a proper understanding of the traded products giving rise to opacity and unpredictability in the market. This potentially increased inefficiency and system failure [92] across the financial system.

Despite criticisms of the self-certification process and its controversial role in the 2007 financial crisis, the CFTC allowed the processing of crypto-derivatives in 2014, as TeraExchange self-certified its bitcoin non-deliverable forwards.[93] Several US futures exchanges such as CME, CBOE, ICE, and ErisX self-certifies both cash-settled and physically-settled cryptocurrency-based (bitcoin and ether) futures contracts and binary options.[94] Although the CFTC stated that the Commission "held rigorous discussions" [95] with the exchanges for weeks before allowing them to self-certify these crypto-derivatives products, such a process spurred agitation

[88] ibid.

[89] For an academic discussion on the CFTC's role in approving complex financial products, see Saule T Omarova, 'Licence to Deal: Mandatory Approval of Complex Financial Products' [2010] 90 Washington University Law Review 63.

[90] Reiners (n 16) 72. The author compiled the data from a publicly available database on the CFTC's website, that clearly indicated a significant increase of the number of new exchange-traded products approved through self-certification process. It also suggested that this might have potentially contributed to the financial crisis of 2007.

[91] Although the economic purpose of the CFTC's self-certification rule was to "reduce the potential threat of market manipulation or congestion", during the financial crisis of 2008, the market could not necessarily extricate themselves from the underlying cash markets and the policy goal of preventing potential harm to such markets from excessive financial speculation.

[92] Steven L. Schwartz, 'Regulating Complexity in Financial Markets' [2010] 87 Washington University Law Review 211.

[93] Casey (n 6).

[94] David Felsenthal and others, 'Clifford Chance Discusses the Role of the CFTC in the Regulation of Bitcoin' (The CLS Blue Sky Blog, 16 February 2018) <http://clsbluesky.law.columbia.edu/2018/02/16/clifford-chance-discusses-the-role-of-the-cftc-in-the-regulation-of-bitcoin/.> accessed 13 November 2020.

[95] Commodity Futures Trading Commission, 'CFTC Statement on Self-Certification of Bitcoin Products by CME, CFE, and Cantor Exchange' (Commodity Futures Trading Commission, 1 December 2017) <www.cftc.gov/PressRoom/PressReleases/pr7654-17> accessed 13 November 2020.

within the futures industry.[96] In 2017, Walk Lukken, the CEO of the Futures Industry Association (FIA), expressed the FIA's concerns:

> "We remain apprehensive with the lack of transparency and regulation of the underlying reference products on which these products are based and whether exchanges have the proper oversight to ensure the reference products are not susceptible to manipulation, fraud, and operational risk".[97]

Besides, the CFTC does not oversee the cryptocurrency spot market, making it more susceptible to fraud and price manipulation.[98] To respond to the FIA's concern and provide more clarity, the CFTC came up with a stricter review for self-certified crypto-derivatives products, that is "heightened review" [99] for all bitcoin futures and crypto-derivatives products that will apply through self-certification.[100] However, such a review is also questionable because the "heightened review" does not provide the CFTC with an effective oversight mechanism for the cryptocurrency spot market, and therefore, cannot minimise the risks of crypto-derivatives (analysis set out below).

B. 'Heightened review' does not provide the CFTC with the desired visibility into the cryptocurrency spot markets

Cryptocurrency spot markets operate in an unregulated space, or with little regulatory clarity. For instance, ErisX, that offers Ethereum-based futures contract, insists that the ErisX spot market is beyond the CFTC's regulatory purview. It says that,

> "[t]he CFTC does not have regulatory oversight over virtual currency products including spot market trading of virtual currencies. ErisX spot market is not licensed, approved, or

[96] Walt Lukken, 'Open Letter to CFTC Chairman Giancarlo Regarding the Listing of Cryptocurrency Derivatives' (The Futures Industry Association, 7 December 2017) <https://fia.org/articles/open-letter-cftc-chairman-giancarlo-regarding-listing-cryptocurrency-derivatives> accessed 24 November 2021.

[97] ibid.

[98] ibid.

[99] Reiners (n 16) 74. "Heightened review is a new process, without statutory process, without statutory basis, that the CFTC is using to review new virtual currency derivative products".

[100] Commodity Futures Trading Commission, 'CFTC Backgrounder on Oversight of and Approach to Virtual Currency Futures' (Commodity Futures Trading Commission, 4 January 2018) <https://www.cftc.gov/sites/default/files/idc/groups/public/@newsroom/documents/file/backgrounder_virtualcurrency01.pdf> accessed 24 February 2021.

> registered with the CFTC and transaction on the *ErisX Spot Market are not subject to CFTC rules, regulations or regulatory oversight* (emphasis added)". [101]

Under the CEA, the CFTC is mandated to prevent market manipulation in the derivatives market, which gives the CFTC an authority to act against the price manipulation of any underlying commodity.[102] To achieve this goal, the CFTC intends to ensure that the self-certified crypto-derivatives contracts are not "readily susceptible to manipulation".[103] The 'heightened review' also allows the CFTC to implement risk-mitigation and oversight mechanisms through heightened margin requirements and information-sharing agreements between cryptocurrency exchanges.[104] The CFTC views that the information-sharing agreements between cryptocurrency exchanges will ensure the CFTC's access to data, that could "facilitate the detection and pursuit of bad actors in the underlying spot market".[105] However, unlike other traditional commodity spot markets, there is no existing US law providing "direct, comprehensive federal oversight of underlying bitcoin or virtual currency spot markets".[106] Many of the platforms are located offshore and are not registered with the CFTC or the SEC. Therefore, the CFTC's satisfaction that the information-sharing agreements would ensure their visibility into the cryptocurrency spot market, is debatable.[107]

Also, in cash-settled cryptocurrency futures, the ability to manipulate depends on "how easily the reference rate that is used to price the contract can

[101] ErisX <www.erisx.com/about/investors/> accessed 24 February 2021.

[102] Sections 6(c)(1), 6(c) (3), and 9(1) Commodity Exchange Act; Commodity Futures Trading Commission, 'Prohibition on the Employment, or Attempted Employment, of Manipulative and Deceptive Devices and Prohibition on Price Manipulation' (Commodity Futures Trading Commission,14 July 2011) (Final Rule 180.1) <www.federalregister.gov/documents/2011/07/14/2011-17549/prohibition-on-the-employment-or-attempted-employment-of-manipulative-and-deceptive-devices-and> accessed 13 November 2020.

[103] Section 38.200, Title 17, Code of Federal Regulations.

[104] Heightened review includes, among others, DCMs to enter into "direct or indirect information sharing agreements with spot market platforms to allow access to trade and trader data". See Commodity Futures Trading Commission (n 100) 3.

[105] Jay Clayton and Cristopher Giancarlo, 'Regulators are Looking at Cryptocurrency' (The Wall Street Journal, 24 January 2018) < https://www.wsj.com/articles/regulators-are-looking-at-cryptocurrency-1516836363> accessed 11 May 2021.

[106] Felsenthal and others (n 94). See also, Jerry Brito, Houman B Shadab, and Andrea Castillo, 'Bitcoin Financial Regulations: Securities, Derivatives, Prediction Markets, and Gambling' [2014] 51 Columbia Science and Technology Law Review 144, 196, arguing "physically-settled transactions are generally not subject to the full scope of CFTC regulation precisely because they do not implicate the markets that the CFTC is concerned about, namely, futures and swaps markets".

[107] Reiners (n 16) 75.

also be manipulated".[108] Although there is no evidence of price manipulation of the CME or ICE futures until April 2021, such manipulation on cryptocurrency exchanges is rampant.[109] In the absence of federal oversight, crypto-exchanges are widely involved in price tampering by creating fake trade volumes (commonly known as "pump-and-dump").[110] In many instances, traders use social media to perform pump-and-dump schemes to inflate the virtual currencies' prices artificially.[111] A recent study has revealed Telegram and Discord's large-scale pump-and-dump scheme.[112] Such pumping-and-dumping activity can hurt investors in the crypto-derivatives markets in the long run because crypto-derivatives enable institutional investors (and also potential manipulators) to bet on the future bitcoin's price. It is not unlikely that a group of traders would place a massive trade on a bitcoin spot market on the contract's settlement date, thereby pushing up the price of bitcoin and earning a profit on the futures position, and in the same way, if the speculation is on a decreased price, instantly dump the trade.

Despite the apprehension, there is a regulatory vacuum in enforcing cryptocurrency pump-and-dump.[113] Usually, the SEC enforces against pump-and-

[108] ibid.

[109] ibid.

[110] Paul Vigna, 'Most Bitcoin Trading Faked by Unregulated Exchanges: Study Finds' (The Wall Street Journal, 22 March 2019) <www.wsj.com/articles/most-bitcoin-trading-faked-by-unregulated-exchanges-study-finds-11553259600?mod=hp_lead_pos7> accessed 13 November 2020. See also, Kate Rooney, 'Majority of Bitcoin Trading is a Hoax' (CNBC, 23 March 2019) <www.cnbc.com/2019/03/22/majority-of-bitcoin-trading-is-a-hoax-new-study-finds.html> accessed 13 November 2020.

[111] JT Hamerick and others, 'An Examination of the Cryptocurrency Pump and Dump Ecosystem' [2021] 58(4) Information Processing and Management 102506 <https://doi.org/10.1016/j.ipm.2021.102506> accessed 11 May 2021.

[112] Michael Mckee, 'Trader Using "Pump and Dump" Schemes to Manipulate Cryptocurrency Prices', (Finbrief, 22 August 2018) <https://blogs.dlapiper.com/globalfinance/2018/08/22/traders-using-pump-and-dump-schemes-to-manipulate-cryptocurrency-prices/> accessed 13 November 2020 ("[s]uch pump and dump scheme are accomplished through private chatrooms which are accessible only by invitation, and generally overseen by an anonymous moderator. The strategy is to announce a date, time, and exchange for a pump of typically illiquid cryptocurrency. As the buying frenzy pushes the prices up, the members of the pump group begin dumping, i.e., selling at the signal. Successful traders gloat about their profits").

[113] A report published by the New York Office of the Attorney General admitted that the regulators lack control to evade pump and dump activity. See New York State Office of the Attorney General, 'Virtual Markets Integrity Initiative Report' (New York State Office of the Attorney General, September 2018) <https://ag.ny.gov./sites/default/files/vmii_report.pdf> accessed 23 November 2020.

dump schemes since it is a type of securities fraud.[114] Nevertheless, in the existing federal securities law, the SEC is not likely to intervene as it has determined that bitcoin is not a security.[115] Hence, the question remains, will the CFTC oversee such illegal activities in the crypto-derivatives commodity spot market under its mandate of preventing fraud and market manipulation? In 2018, the CFTC first issued an advisory note to warn the consumers to beware and avoid pump-and-dump schemes that occurred in cryptocurrency trading.[116] Nonetheless, the CFTC's overall approach does not adequately address the fraud and manipulation concerns in the cryptocurrency spot market.[117] On the other hand, in the absence of any oversight mechanisms, cryptocurrency spot markets operate in an unregulated space, that is likely to incentivise manipulative behaviour in the crypto-derivatives market.

C. Crypto-derivatives serve only the interest of cryptocurrency exchanges

Many scholars argued that crypto-derivatives would serve only the interest of the cryptocurrency exchanges because crypto-derivatives allow these exchanges to hedge their risk exposures that arise from the volatility in the cryptocurrency spot market.[118] To illustrate, if someone purchases a piece of furniture on Overstock and pays in bitcoin via Coinbase, the payment is denominated in Dollars and transferred from Coinbase to Overstock's bank account. This means "it is Coinbase that is accepting the exchange volatility risk".[119] Even though Coinbase charges Overstock a certain percentage as a payment processing fee, such a fee is not sufficient to "cover the exchange rate risk that Coinbase could

[114] Section 10(b) read with Section 17(a) (2) and Rule 10b-5, Securities Exchange Act 1934. For a brief analysis, see Wendy Gerwick Couture, 'Prosecuting Securities Fraud under Section 17 (a) (2)' (The CLS Blue Sky Blog, 20 March 2019) < https://clsbluesky.law.columbia.edu/2019/03/20/prosecuting-securities-fraud-under-section-17a2/> accessed 11 May 2021.

[115] Eugene Kim, 'The SEC Warns Investors About Potential ICO Scams and 'Pump and Dump' Schemes' (CNBC, 28 August 2017) <www.cnbc.com/2017/08/28/sec-warns-on-ico-scams-pump-and-dump-schemes.html> accessed 13 November 2020.

[116] Commodity Futures Trading Commission, 'CFTC Warns Customers to Avoid Pump-and-Dump Schemes' (Commodity Futures Trading Commission, 15 February 2018) < www.cftc.gov/PressRoom/PressReleases/pr7697-18> accessed 22 November 2020.

[117] In the past, the CFTC publicly announced that it is not the CFTC's duty to oversee a spot market on a daily basis. See Reiner (n 16) 85.

[118] Brito, Shadab, and Castillo (n 107).

[119] Cade Metz, 'The Grand Experiment Goes Live: Overstock.com is Now Accepting Bitcoin' (The Wired, 4 January 2014) <www.wired.com/2014/01/overstock-bitcoin-live/> (discussing the volatility risks associated with Bitcoin and Overstock's collaboration with Coinbase).

face in the future".[120] It will only make sense if Coinbase can hedge its exchange rate risk by "simply engaging in swap or futures contract".[121] Arguably, this is one of the reasons why cryptocurrency exchanges were insisting on crypto-derivatives for so long.[122] Also, fraud,[123] scams,[124] hacks,[125] and insider trading,[126] are rampant in the cryptocurrency market. Moreover, it is quite uncertain what impacts on the market price of cryptocurrencies. Any regulatory move appears to impact bitcoin price; bitcoin price dropped by 30% when China banned cryptocurrency or South Korea initiated a crackdown on cryptocurrency.[127] In March 2020, when the US market was turbulent due to the COVID-19 crisis combined with a plummet in oil prices and sell-off in stocks,[128] the cryptocurrency market lost almost $26.43 billion in a day (see Figure III.1).[129] Since November 2020, bitcoin's price has been soaring and hit as high as $63,729.50.[130] In addition to spot markets, the abrupt price dips

[120] Brito, Shadab, and Castillo (n 107) 157.

[121] ibid 157–58.

[122] ibid.

[123] Collen Shalby, 'Camarillo Man and Two Others Arrested in Alleged $722-million Cryptocurrency Fraud Scheme' (The Los Angeles Times, 10 December 2019) <www.latimes.com/california/story/2019-12-10/camarillo-man-and-two-others-arrested-in-alleged-722-million-cryptocurrency-fraud-scheme> accessed 13 November 2020.

[124] Shaurya Malwa, 'Twitter Bitcoin Scams Take New Leap After Verified Twitter Accounts Impersonate Elon Musk' (CryptoSlate, 18 November 2019) <https://cryptoslate.com/twitter-bitcoin-scams-take-new-leap-after-verified-twitter-accounts-impersonate-elon-musk/> accessed 13 November 2020.

[125] Eric Lam, 'Hackers Steal $40 Million Worth of Bitcoin from Binance Exchange' (The Bloomberg, 08 May 2019) <www.bloomberg.com/news/articles/2019-05-08/crypto-exchange-giant-binance-reports-a-hack-of-7-000-bitcoin> accessed 21 November 2020; Andrew Norry, 'The History of the Mt. Gox Hack: Bitcoin's Biggest Hit' (Blockonomi, 7 June 2019) <https://blockonomi.com/mt-gox-hack/> accessed 13 November 2020.

[126] Daniel Oberhaus, 'Coinbase is Being Sued for Insider Trading' (Vice, 5 March 2018) <www.vice.com/en_us/article/pam4xn/coinbase-insider-trading-lawsuit-gdax-bitcoin-cash> accessed 13 November 2020.

[127] Stefan Stankovic, 'US Cryptocurrency Regulation: Policies, Regimes & More' (Unblock, 18 February 2019) <https://unblock.net/us-cryptocurrency-regulation/#h3> accessed 21 November 2020.

[128] Luke Kawa, 'Stock Market Volatility Tops Financial Crisis with VIX at Record' (Bloomberg, 16 March 2020) <https://www.bloomberg.com/news/articles/2020-03-16/stock-market-volatility-tops-financial-crisis-with-vix-at-record> accessed 11 May 2021.

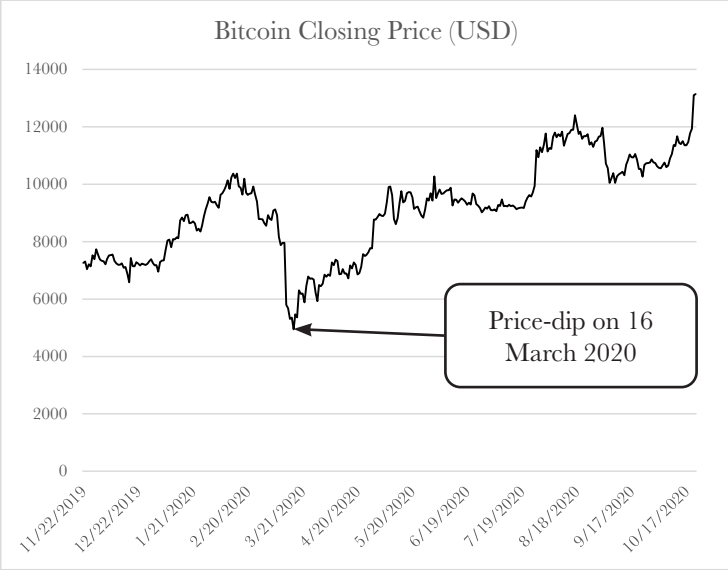
[129] Arjun Kharppal, 'Over $26 Billion Wiped Off Cryptocurrency Market in 24 Hours After Massive Oil Price Plunge' (CNBC, 8 March 2020) <www.cnbc.com/2020/03/09/bitcoin-btc-and-other-cryptocurrency-prices-plunge-after-oil-drop.html?__source=sharebar|linkedin&par=sharebar> accessed 21 November 2020.

[130] Ryan Browne, 'Bitcoin Hits New All-Time High above $63,000 ahead of Coinbase Debut' (CNBC, 13 April 2021) <https://www.cnbc.com/2021/04/13/bitcoin-hits-new-all-time-high-above-62000-ahead-of-coinbase-debut.html> accessed 11 May 2021.

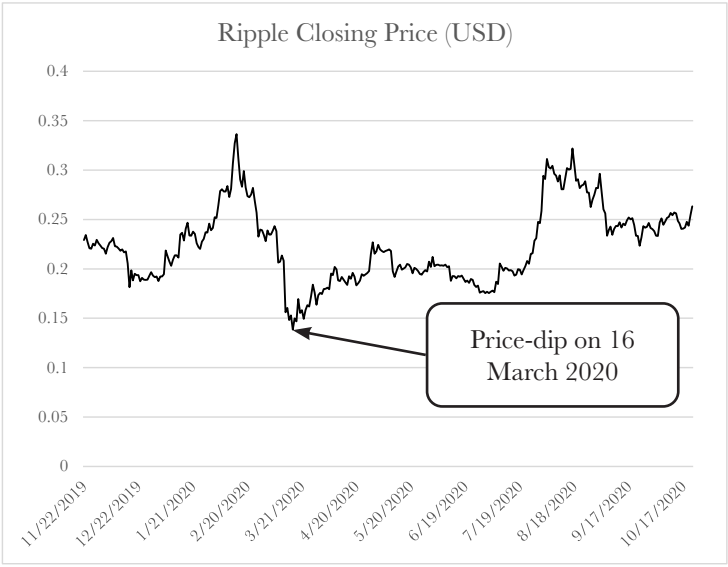
regularly wipe out billions of dollars from other cryptocurrency-based markets, such as the decentralised finance market.[131]

FIGURE III.1

Daily Closing Prices of Bitcoin, Ethereum, and Ripple (22 November 2019–21 November 2020)



[131] Jose Antonio Lanz, 'Ethereum Price Dip Wipes $1.5 Billion from DeFi Markets' (Decrypt, 31 October 2020) <https://decrypt.co/46796/ethereum-price-1-5-billion-defi-markets> accessed 21 November 2020.

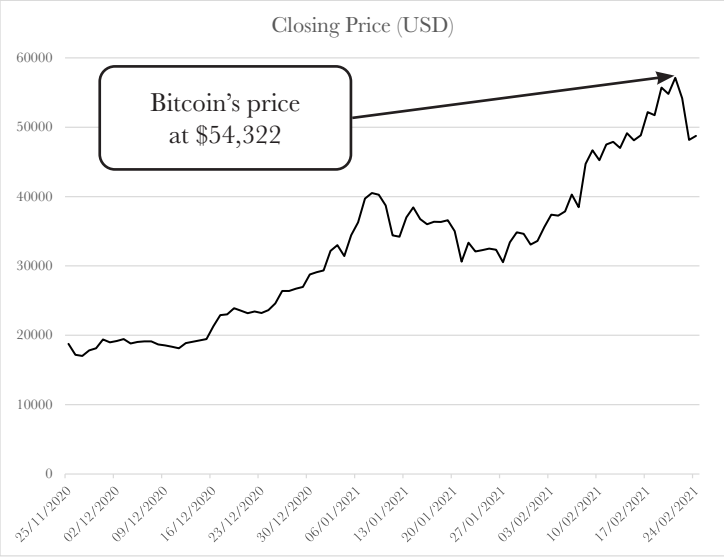


Source: Coindesk

Figure III.1 shows that all three cryptocurrencies suffered a major price dip on 16 March 2020 as a response to the COVID-19 crisis coupled with the turmoil in the US financial markets and international oil prices.

FIGURE III.2

Bitcoin's Soaring Price between (25 November 2020–24 February 2021)



Source: Coindesk

In light of cryptocurrency's price volatility and regulatory opacity, the addition of federally regulated crypto-derivatives to the market will give these investors complacency that their investments are protected under federal laws. However, it appears that until now the regulators' approach is largely based on warnings where customers are advised to do their research before nvesting in cryptocurrency-related products. issued from time to time.[132] Therefore, in the

[132] Commodity Futures Trading Commission, 'Investor Alert, Watch Out for Fraudulent Digital Asset and "Crypto" Trading Websites' (26 April 2019) <https://www.cftc.gov/sites/default/files/2019-04/OIEA%20and%20CFTC%20Investor%20Alert%20Fraudulent%20Digital%20Assets%20Websites.pdf> accessed 11 May 2021.

context of an unregulated spot market and the CFTC's lack of oversight on it, investor protection in the crypto-derivatives market is very much questionable.

D. The SEC contradicts the CFTC's view as the SEC rejects ETFs on the ground of fraud and abusive practices in the cryptocurrency spot market

Unlike the CFTC, the SEC appears to be reluctant to approve any new cryptocurrency products that would require oversight in the spot market.[133] On several occasions, SEC has raised concerns about the current cryptocurrency markets featuring less investor protection and more susceptibility to fraud and manipulation.[134] The SEC's view became evident when it first rejected a bitcoin exchange-traded product ('ETP')[135] in 2018.[136] To date, the SEC has disapproved of more than nine bitcoin ETP proposals, including bitcoin ETP proposals from

[133] See Marion A. Brown, 'Cryptocurrency and Financial Regulation: The SEC's Rejection of Bitcoin-Based ETPs' [2012] 23 N.C. Banking Institute 139. See also, Tom Lydon, 'SEC Rejects 9 Applications for Bitcoin ETFs' (The NASDAQ, 23 August 2018) <www.nasdaq.com/articles/sec-rejects-9-applications-bitcoin-etfs-2018-08-23> accessed 13 November 2020.

[134] See Jay Clayton, 'Statement on Cryptocurrencies and Initial Coin Offerings' (US Securities and Exchange Commission, 11 December 2017) <www.sec.gov/news/public-statement/statement-clayton-2017-12-11> accessed 14 November 2020 (warning that cryptocurrency markers span national borders as the investment funds may travel across boundaries rapidly. As a result, "risks can be amplified, including the risk that market regulators, such as the SEC, may not be able to effectively pursue bad actors or recover funds"). See also, Jay Clayton and Christopher Giancarlo, 'Statement by SEC Chairman Jay Clayton and CFTC Chairman J. Christopher Giancarlo: Regulators are Looking at Cryptocurrency' (US Securities and Exchange Commission, 25 January 2018) <www.sec.gov/news/public-statement/statement-clayton-giancarlo-012518> accessed 13 November 2020.

[135] ETPs are securities that are traded on exchanges similar to stocks. Cryptocurrency ETPs could be in two basic forms: (1) ETFs holding crypto-derivatives; and (2) ETPs physically holding cryptocurrency. See also, James Chan, 'Exchange Traded Products (ETPs)' (Investopedia, 25 August 2019) <www.investopedia.com/terms/e/exchange-traded-products-etp.asp> accessed 13 November 2020.

[136] Kate Rooney and Bob Pisani, 'Winklevoss Twins Bitcoin ETF Rejected by SEC' (CNBC, 26 July 2018) <www.cnbc.com/2018/07/26/winklevoss-twins-bitcoin-etf-rejected-by-sec.html> accessed 13 November 2020.

Proshares,[137] Direxion,[138] and GraniteShares,[139] and the latest being the ETP application by Wilshire Phoenix.[140] In rejecting the bitcoin ETPs applications, the SEC held that none of the applicants had proved that the cryptocurrency market is uniquely resistant to market manipulation to secure investor protection and public interest, as required under §6(b)(f) of the Exchange Act 1934. These ETP products are not designed to "prevent fraudulent and manipulative acts and practices",[141] as the current bitcoin futures market (such as CMR and CBOE) is not of a significant size. This technically prevents the DCMs from detecting and deterring misconduct and tracing price manipulation despite the use of an information-sharing agreement.[142] With regard to investor protection, although the SEC viewed that "trading a bitcoin-based ETP on a national securities exchange might provide some additional protection to investors",[143] this protection is not sufficient to fulfil the requirements of §6(b)(f) of the Exchange Act, that requires the "rules of a national securities exchange be designed to prevent fraudulent and manipulative acts and practices".[144]

Regarding these new cryptocurrency products, the SEC also raises the issue of custody risk.[145] As bitcoin is largely traded on unregulated international exchanges, these custodians carry a significant risk of being hacked or going out

[137] Securities and Exchange Commission, 'SEC Release No. 82350' (19 December 2017); Securities and Exchange Commission, 'SEC Release No. 82 FR 61100', 26 December 2017).

[138] Securities and Exchange Commission, 'SEC Release Nos. 82532' (18 January 2018); and Securities and Exchange Commission, 83 FR 3380 (SR-NYSEArca-2018-02, 24 January 2018).

[139] Securities and Exchange Commission, 'SEC Release No. 34-83913' (22 August 2018) <www.sec.gov/rules/sro/cboebzx/2018/34-83913.pdf> accessed 13 November 2020. For commentary see Nikhilesh De, 'The Securities and Exchange Commission (SEC) has Issued Rejections to Bitcoin Exchange-Traded Fund (ETFs) Proposals from ProShares, Direxion and GraniteShares' (Coindesk, 27 November 2019) <www.coindesk.com/sec-rejects-7-bitcoin-etf-proposals> accessed 13 November 2020.

[140] Nikhilesh De, 'SEC Rejects Latest Bitcoin ETF Bid' (Coindesk, 27 February 2020) <www.coindesk.com/sec-rejects-latest-bitcoin-etf-bid> accessed 14 November 2020.

[141] See Katie Rooney and Bob Pisani, 'Winklevoss Twins Bitcoin ETF Rejected by SEC' (CNBC, 26 July 2018) <www.cnbc.com/2018/07/26/winklevoss-twins-bitcoin-etf-rejected-by-sec.html> accessed 30 November 2019).

[142] Securities and Exchange Commission (n 139) 24. The rationale here is that to successfully manipulate the ETP, one would also have to trade on the spot market. In traditional commodity a surveillance-sharing agreement assists the ETP listing markets in spotting manipulative behaviour in the spot market.

[143] ibid 29.

[144] ibid.

[145] ibid.

of business.[146] Custody risk is also present in the crypto-derivatives markets, which is rarely addressed by the CFTC. Particularly, for physically-settled bitcoin futures contracts, the exchanges are required to hold physical bitcoins.[147] Given that there is no federal-level investor protection for these trusts, bitcoin held by the trust is not subject to Federal Deposit Insurance Corporation or SIPA. Therefore, if cryptocurrencies are lost or stolen, or a crypto holder dies, it is likely that those coins will be lost forever.[148] The existing law hardly "adjudicate the matter of recovering the coins owned by the crypto holder who has passed away".[149]

E. The cftc's approach deviates from the major global regulators

The CFTC's approach in approving crypto-derivatives deviates from the other two major global regulators, the ESMA and the FCA. Both the EU and the UK regulators took measures restricting the trading of crypto-derivatives, determining that retail investors are not protected from the price volatility, speculation, and other forms of market and operational risks associated with cryptocurrencies.[150] Also, the complexity of these products and a lack of transparency limit retail investors' ability to understand the risks underlying these products.[151] Leveraged crypto-derivatives are risky and extremely volatile, increasing the scale and speed

[146] Daniel Shane, 'Bitcoin Exchange Goes Bust After Hack' (CNN, 20 December 2017) <https://money.cnn.com/2017/12/20/technology/south-korea-bitcoin-exchange-closes/index.html> accessed 23 November 2020.

[147] A company's CEO has gone missing with Cold Wallet's access. See Matthew Beedham, 'Cryptocurrency Exchange IDAX's CEO Reportedly Missing with Company's Wallet' (The Next Web, 29 November 2019) <https://thenextweb.com/hardfork/2019/11/29/cryptocurrency-exchange-ceo-missing-idax-bitcoin-cold-wallet/> accessed 14 November 2020. On another occasion, a CEO died and was the only person who knew the password of the company's cold wallet. See Antonia Noor Farzan, 'Millions Vanished with a Cryptocurrency Entrepreneur's Sudden Death. Now Investors Want His Body Exhumed' (The Washington Post, 16 December 2019) <www.washingtonpost.com/nation/2019/12/16/gerald-cotten-quadrigacx-cryptocurrency-death-body-exhumed/> accessed 14 November 2020. In both cases, the investors lost all the cryptocurrencies and the money.

[148] Lefan Gong and Luping Yu, 'China' in Josias Dewey (ed), Blockchain & Cryptocurrency Regulation (Global Legal Group 2019) 261, 266 <www.acc.com/sites/default/files/resources/vl/membersonly/Article/1489775_1.pdf> accessed 24 February 2021.

[149] ibid.

[150] See analyses in Section II.A and II.B regarding the EU and UK's ban on crypto-derivatives on the ground of potential harm to retail investors.

[151] Financial Conduct Authority, 'Prohibiting the Sale to Retail Clients of Investment Products that Reference Cryptoassets' (n 2); Financial Conduct Authority, 'Prohibiting the Sale to Retail Client of Investment Products that Reference Cryptoassets' (n 54).

of investors' losses from a crypto-derivative.[152] Furthermore, the regulators should be cautious about the risk of the cryptocurrency speculative bubble. Although many crypto-enthusiasts believe that crypto-derivatives could increase liquidity in the cryptocurrency market, thereby stabilizing price volatility,[153] in the absence of a comprehensive regulation addressing the core regulatory concerns, crypto-derivatives would still hurt retail investors.

IV. Future of Crypto-Derivatives in the US: Possible Regulatory Frameworks

To achieve a robust and effective crypto-derivatives regulatory framework in the US, it is essential that: (1) the US crypto-derivatives market is free from manipulative and abusive practices; (2) regulators have adequate visibility into the cryptocurrency spot market; (3) regulators are well-equipped to detect abusive and manipulative practices in the crypto-derivatives market and; (4) enforcement mechanisms are in place to safeguard investors' interest. This paper explores two possible regulatory frameworks for crypto-derivatives. First, like the UK (and possibly the EU in the future), there could be a complete ban on crypto-derivatives enacted by a federal statute, as crypto-derivatives are just a means of speculation, and the lack of oversight in the spot market will continue harming retail investors. Second, if an outright ban is not feasible, Congress must develop comprehensive legislation that recognises the novel market and operational risks posed by cryptocurrency, and introduce effective regulatory intervention in the crypto-derivatives markets.

A. The possibility of a complete ban on crypto-derivatives

Following the UK and the EU, the US regulators may consider banning the sale and purchase of crypto-derivatives in the derivatives exchanges, keeping investor protection as their central focus. Crypto-derivatives pose a unique threat to investors due to their high leverage and extreme price volatility.[154] Moreover, the failure of the CFTC to have any oversight mechanism on the cryptocurrency

[152] ibid.

[153] Many were of the view that having another competitor in the market or other altcoin derivatives could give a major boost to its liquidity and trading volumes. This may also create awareness of broader cryptocurrency market among investors that may result in infusing more money into the market. This could help create less volatility in altcoin prices.

[154] Ryan Clements, 'Cryptocurrency Self-Regulatory Organization (CSRO)' (The FinReg Blog, 21 June 2019) <https://sites.duke.edu/thefinregblog/2018/06/21/can-a-cryptocurrency-self-regulatory-organization-work-assessing-its-promise-and-likely-challenges/> accessed 23 November 2020.

spot market is contrary to the CFTC's mandate to protect investors from fraud and abusive market practices. Unless the CFTC has established a mechanism to oversee the cryptocurrency spot market meaningfully and adequately, a ban on crypto-derivatives will act as a warning to investors not to put their money in such risky products. Further, the complexity of the crypto-derivatives and the lack of transparency in the cryptocurrency spot market require a more rigorous enforcement approach from the CFTC. However, many argue that an outright ban is likely to hurt the existing cryptocurrency platforms that comply with the laws and regulations.[155] There is another set of arguments:

> "where regulator erred on the side of banning or bashing cryptocurrencies, they have faced classical problems of regulatory competition and regulatory arbitrage, i.e., the migration of the industry from their jurisdiction to more welcoming ones or migration of activities to underground or black-markets".[156]

Given the drawbacks of an outright ban on crypto-derivatives, this paper proposes an alternative regulatory framework — enacting comprehensive federal-level crypto-regulation in response to the emerging issues of manipulation and lack of investor protection in the cryptocurrency spot market and vis-à-vis crypto-derivatives markets.

B. The need for a comprehensive federal crypto-regulation

The need for a comprehensive crypto-regulation is premised on four grounds. First, the sporadic regulatory efforts among different US regulatory agencies concerning cryptocurrency are counterproductive. A systemic regulatory approach can minimise the risks of cryptocurrency spot markets and avert market failure.[157]

Second, the novelty involved in cryptocurrency requires a uniform regulatory approach. Otherwise, it may bring about an unwanted disruption in the capital and financial market. In the US, the regulatory approach to cryptocurrency is fragmented. For instance, while the SEC has declared ICO as a security, it did not establish its jurisdiction exclusively on all digital tokens. Meanwhile, the US Internal Revenue Service (IRS) considers convertible cryptocurrency as property

[155] Osato Avan-Nomayo 'Cryptoderivatives Ban is Not a Good Idea, Says WFE' (Blockonomi, 08 October 2019), <https://blockonomi.com/crypto-derivatives-ban-not-good-idea-says-wfe/> accessed 23 November 2020.

[156] Hossein Nabilou, 'How to Regulate Bitcoin? Decentralized Regulation for a Decentralized Cryptocurrency' [2019] 27 International Journal on Law and Information Technology 266, 270.

[157] ibid.

for tax purposes.[158] In addition, the federal courts in several cases (such as *United States v. Ulbricht*)[159] have treated cryptocurrency as money for specific purposes. However, there are still divided opinions as to the status of cryptocurrency as "money" because: (1) cryptocurrency is not widely accepted as a means of payment and (2) its store value is unreliable due to market volatility.[160] It is argued that a comprehensive federal crypto-regulation can cure the problem of this fragmented, sporadic, and ambiguous regulatory approach by adopting a unanimous definition of cryptocurrency (or by defining it as a separate 'digital asset class') and thus bringing the intermediaries and cryptocurrency-based assets under the same regulatory framework.[161]

Third, if fraud and market manipulation continue, it will eventually drive potential investors away from the market. Furthermore, opacity and lack of regulatory clarity can result in the loss of investors' confidence.[162]

Finally, if cryptocurrencies become an effective monetary instrument in the future, its impact on the country's monetary policy would be profound as the Federal Reserve Board might lose its ability to control the money supply.[163] Therefore, there is a demand from both policymakers and market participants that

158 Internal Revenue Service, 'Virtual Currencies' (The Internal Revenue Service) <www.irs.gov/businesses/small-businesses-self-employed/virtual-currencies> accessed 21 November 2020.

159 United States of America v. Ulbricht, No. 15-1815 (2d Cir 2017). Retrieved from <https://cases.justia.com/federal/appellate-courts/ca2/15-1815/15-1815-2017-05-31.pdf?ts=1496241010> accessed 11 May 2021.

160 Mohamed Damak, 'The Future of Banking: Cryptocurrencies Will Need Some Rules to Change the Game' (S&P Global, 19 February 2018) <www.spglobal.com/en/research-insights/articles/the-future-of-banking-cryptocurrencies-will-need-some-rules-to-change-the-game> accessed 4 December 2019. As it goes "[…] cryptocurrencies do not meet the basic two requisites of a currency: An effective mean of exchange and an effective store of value. First, cryptocurrencies are still not widely accepted as payment instruments, although the list of companies accepting them have increased over the past few years. Second, the volatility that we have observed over the past 12 months in the valuation of some cryptocurrencies and their market cap is the most meaningful evidence that they fail the test of value storage. We also don't view cryptocurrencies as an asset class. For starters, the total outstanding aren't big enough yet. At Feb. 10, 2018, there were 1,523 outstanding cryptocurrencies with a market cap of around $394 billion. By way of comparison, at the same date, this is well below the market capitalization of Apple Inc., around $794 billion".

161 Averie Brooks, 'U.S. Regulation of Blockchain Currencies: A Policy Overview' [2018] 9 Amherst Intellectual Property Brief 75.

162 ibid.

163 Damak (n 160).

Congress develop comprehensive federal legislation to regulate cryptocurrencies[164] and bring it under a uniform legislative scope.[165]

In the UK and the EU, regulators are taking measures to establish a uniform and robust regulatory framework to achieve transparency into the cryptocurrency spot market. The UK government is in the consultative process with various stakeholders and industry participants to ensure that "its regulatory framework is equipped to harness the benefits of new technologies, supporting innovation and competition, while mitigating risks to consumers and stability".[166] The UK's efforts on regulatory measures predominantly aim to enhance consumer protection and address risks and challenges associated with cryptocurrency and stable coins.[167] With a similar approach, the European Commission has adopted a comprehensive digital finance package that included "Legislative Proposals on Cryptoassets",[168] to provide the cryptocurrency markets with coherent legal rules as well as to support financial innovation, reinforce investor protection while ensuring financial stability.[169] The proposal further aims to reduce the market fragmentation by

[164] Financial Stability Oversight Council, 'Annual Report' (2019) <https://home.treasury.gov/system/files/261/FSOC2019AnnualReport.pdf> accessed 20 November 2020 (calling for stricter state and federal regulations of stablecoins and digital assets).

[165] Peter Van Valkenburgh, 'A National Alternative to Onerous State-by-State Regulation of Cryptocurrency Intermediaries' (Coin Center, 30 August 2019) <https://coincenter.org/entry/a-national-alternative-to-onerous-state-by-state-regulation-of-cryptocurrency-intermediaries> accessed 20 November 2020.

[166] See HM Treasury, 'UK Regulatory Approach to Cryptoassets and Stablecoins: Consultation and Call for Evidence' (January 2021) 3 <www.gov.uk/government/consultations/uk-regulatory-approach-to-cryptoassets-and-stablecoins-consultation-and-call-for-evidence> accessed 21 February 2021.

[167] ibid. The UK's Call for Evidence report is the reflection of the final report submitted by the Cryptoassets Taskforce in 2018 which advised the government to take actions in five main grounds to: (1) "maintain the UK's international reputation as a safe and transparent place to do business in financial services; (2) ensure high regulatory standards in financial markets; (3) protect consumers; (4) guard against threats to financial stability that could emerge in the future; and (5) allow those investors in the financial sector that play by the rules to thrive". See also, HM Treasury, Financial Conduct Authority, and Bank of England, 'Cryptoassets Taskforce: Final Report' (October 2018) 6 < https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/752070/cryptoassets_taskforce_final_report_final_web.pdf> accessed 21 February 2021.

[168] The proposed legislative proposal on cryptoassets will be accompanied by the MiFID and MiFIR.

[169] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council on Markets in Crypto-assets, and amending Directive (EU) 2019/1937' [2020] COM/2020/593 final <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX-:52020PC0593> accessed 21 February 2021.

developing "uniform conditions of operations for firms with the EU",[170] that can be utilised to overcome regulatory differences across the member states.

In contrast, the US regulatory approach towards cryptocurrency is still sporadic, fragmented, and ambiguous, along with powers being divided across multiple regulatory agencies between SEC, CFTC, the Office of the Comptroller of the Currency ('OCC'), the Financial Crimes Enforcement Network ('FinCEN') and Internal Service Revenue ('IRS').[171] In addition to the SEC and the CFTC's authority over cryptocurrency, the OCC from time to time, provides interpretative letters and guidance for the banks and financial institutions to delineate the permissible activities concerning cryptocurrency.[172] Since 2013, the FinCEN has also been issuing instructions for banks, Money Services Businesses (MSBs), and

[170] ibid 5.

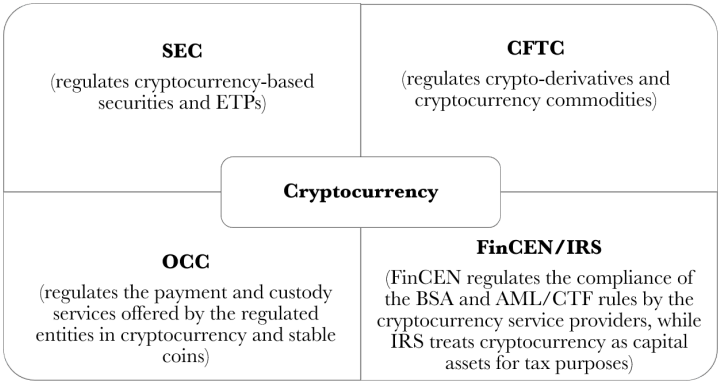
[171] For an analysis on the gap in the regulation of cryptoassets, see Timothy G. Massad, 'It's Time to Strengthen the Regulation of Crypto-Assets' (Brookings, 18 March 2019) <https://www.brookings.edu/wp-content/uploads/2019/03/Timothy-Massad-Its-Time-to-Strengthen-the-Regulation-of-Crypto-Assets-2.pdf> accessed 11 May 2021.

[172] In an interpretive letter dated 04 January 2021, the OCC has now permitted the banks to use cryptocurrency and stable coins to facilitate payment transactions for customers. The banks can also validate, store and record payment transactions made in cryptocurrency and stablecoins, and serve as a node of a blockchain (INVN). See Office of the Comptroller of the Currency, 'OCC Chief Counsel's Interpretation on National Bank and Federal Savings Association Authority to Use Independent Node Verification Networks and Stablecoins for Payment Activities' (04 January 2021) <www.occ.treas.gov/news-issuances/news-releases/2021/nr-occ-2021-2a.pdf> accessed 24 February 2021. In a previous interpretive letter dated 22 July 2020, the OCC permitted the US national banks to provide cryptocurrency custody services for customers. See, Office of the Comptroller of the Currency, 'Authority of a National Bank to Provide Cryptocurrency Custody Services for Customers' (22 July 2020) <www.occ.treas.gov/topics/charters-and-licensing/interpretations-and-actions/2020/int1170.pdf> accessed 24 February 2021.

cryptocurrency exchanges to require them to comply with the Bank Secrecy Act[173] and AML/CTF rules.[174]

FIGURE IV.1

The Major US Regulators with Jurisdictions over Cryptocurrency



However, legislative efforts are going on to bring about some regulatory clarities for cryptocurrency service providers and market participants. In particular, the shift towards contactless digital payment because of the COVID-19 pandemic, and Facebook's efforts to initiate their own digital currency, compelled the US legislatures to consider the need for a cryptocurrency regulation. In March 2020, the US House of Representatives proposed a new Cryptocurrency Act 2020,[175] categorizing cryptocurrency or digital tokens into three main groups based on its decentralised nature and the use of cryptographic ledger, that is: (1) cryptocurrencies

[173] 31 United States Code 5311 et seq.

[174] See Financial Crime Enforcement Network, 'FinCEN Issues Guidance on Virtual Currencies and Regulatory Responsibilities' (Financial Crime Enforcement Network, 18 March 2013) <www.fincen.gov/news/news-releases/fincen-issues-guidance-virtual-currencies-and-regulatory-responsibilities> accessed 24 February 2021; Financial Crime Enforcement Network, 'Application of FinCEN's Regulations to Certain Business Models Involving Convertible Virtual Currencies' (Financial Crime Enforcement Network, 9 May 2019 <www.fincen.gov/sites/default/files/2019-05/FinCEN%20Guidance%20CVC%20FINAL%20508.pdf> accessed 24 February 2021; US Department of the Treasury, 'The Financial Crimes Enforcement Network Proposes Rule Aimed at Closing Anti-Money Laundering Regulatory Gaps for Certain Convertible Virtual Currency and Digital Asset Transactions' (18 December 2020) <https://home.treasury.gov/news/press-releases/sm1216> accessed 24 February 2021.

[175] United States House of Representatives, 'H.R.6154': Crypto-Currency Act of 2020.

including the US currency representation; (2) crypto-commodities residing on a blockchain or decentralised cryptographic ledger and; (3) crypto-securities that meet the Howey test.[176] The Act further proposed that depending on the categories, the CFTC, the SEC, and the FinCEN would have regulatory authorities over this asset class. In another draft bill, the Digital Commodity Exchange Act 2020,[177] The US House of Representatives proposed an amendment to the CEA by incorporating definitions of 'digital commodity,' 'digital commodity custodian'[178] and 'digital commodity exchange'.[179] Although it recommended the CFTC as a single regulatory body with exclusive jurisdiction over cryptocurrency-related transactions,[180] the proposition is based on the assumption that cryptocurrencies are only 'virtual commodities', and thus excluded other cryptocurrency-based financial products, such as ICOs and various forms of digital tokens and utility tokens. It also does not provide any specific regulatory guidance where cryptocurrency is used as a payment method.

Nonetheless, both propositions do not adequately answer the regulatory quandaries regarding cryptocurrency, in as much as they do not complement the existing fragmented regulatory approach towards cryptocurrency. To provide a degree of regulatory clarity, this paper proposes a federal level crypto-regulation with a separate regulatory agency having exclusive jurisdiction over cryptocurrencies in the US, including the spot market and any financial instruments where the underlying asset is a cryptocurrency (such as crypto-derivatives). The legislation should also incorporate mandatory registration requirement for cryptocurrency exchanges.[181] The agency will also coordinate with other regulatory agencies, as the new agency will work parallelly with the others (such as the SEC, CFTC, IRS, and FinCEN), but will only be limited to regulating cryptocurrency and other digital assets. A uniform federal-level cryptocurrency agency is likely to establish

[176] Securities and Exchange Commission (n 18).

[177] United States House of Representatives, 'H.R. 8373': Digital Commodity Exchange Act of 2020.

[178] ibid 2. "The term 'digital commodity custodian' means an entity that holds, maintains, or safeguards digital commodities and other assets on behalf of digital commodity market participants".

[179] ibid 2. "The term 'digital commodity exchange' means a trading facility that lists for one digital commodity".

[180] ibid.

[181] The proposed new legislation should also exempt the SEC and the CFTC from exercising its jurisdiction over cryptocurrency-based financial instruments.

effective oversight and supervisory authority over the cryptocurrency spot markets, that would help curb market manipulation and restore investor confidence.[182]

C. The proposed framework of the crypto-regulation

The proposed crypto-regulation should be based on information, equal access, and investors' confidence. In particular, it should have the mandate of protecting investors against fraud and offer a degree of centralization. Therefore, the regulation should have: (1) a new federal cryptocurrency agency established by an Act of Congress; (2) the mandatory registration requirement for all cryptocurrency exchanges, including the cryptocurrency spot markets and; (3) a national cryptocurrency exchange.

(i) A New Federal Cryptocurrency Agency: A de novo Crypto-Regulatory Regime Established by an Act of Congress

A federal cryptocurrency agency having exclusive jurisdiction over cryptocurrencies can prevent price manipulation, fraud, and abusive market practices, by exercising direct oversight over the cryptocurrency intermediaries, including exchanges and spot markets. Although the new cryptocurrency agency structure is subject to rigorous academic, technical, and regulatory discussions, it is not uncommon in the US to constitute a new federal agency to protect consumers and fill in the regulatory vacuum. After the financial crisis of 2008, the Dodd-Frank Wall Street Reform and Consumer Protection Act established the Consumer Financial Protection Bureau — a single, independent consumer-focused regulatory regime — consolidating the scattered financial authorities throughout the federal government and bringing them under one roof.[183] Similarly, the de novo crypto-regulator should combine existing regulators' mandates, jurisdictions, responsibilities, and enforcement authorities over cryptocurrencies, and overcome the current regulatory overlap and ambiguity. The de novo regime's mandate should be based on investor protection through promoting market transparency.

With respect to jurisdictions, this paper proposes that the crypto-regulator should deal with cryptocurrency-based securities (for example, ICOs, digital tokens, and utilities), derivatives (for example, swaps, futures, and options), ETFs,

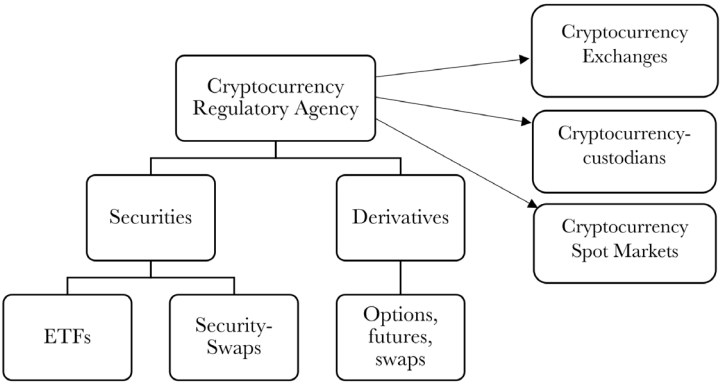
[182] This paper does not propose a direct regulation of the cryptocurrency technology (i.e., blockchain technology) itself. Rather, the regulation should be indirect, meaning it should address the regulation of the cryptocurrency intermediaries, including cryptocurrency exchanges and custodians.

[183] 12 United States Code Sections 5491–5497 (2010).

crypto-securities based swaps, cryptocurrency custodians, and cryptocurrency spot markets (Figure IV.2).

FIGURE IV.2

The Proposed US Crypto-Regulation



With respect to jurisdictions, this paper proposes that the crypto-regulator should deal with cryptocurrency-based securities (for example, ICOs, digital tokens, and utilities), derivatives (for example, swaps, futures, and options), ETFs, crypto-securities based swaps, cryptocurrency custodians, and cryptocurrency spot markets (Figure IV.2).

(ii) Federal Licensing Requirement: Overseeing the Cryptocurrency Spot Market

The new regulator should be responsible for authorising licenses to eligible and trusted cryptocurrency intermediaries, i.e., cryptocurrency exchanges, cryptocurrency custodians, payment processors, and cryptocurrency spot markets, that will operate businesses in cryptocurrencies. The licensing regime is significant for cryptocurrency companies because it will help the regulator bring the licensed entities under a single supervisory authority. For instance, after the infamous Mt. Gox hacking incident,[184] Japan enacted an amendment to the Payment Services Act

[184] In 2014, Bitcoin hacking bankrupted a leading exchange in Japan called Mt. Gox, with approximately half-a-billion dollars in bitcoin ($850,000 bitcoin) stolen. While $200,000 bitcoin were recovered within six months, its dollar value sunk by the revelation of weak security, and the incident showed that hacks impact bitcoin's trading price. Exchange customers had no remedy. See Robert McMillan, 'The Inside Story of Mt. Gox, Bitcoin's $460 Million Disaster' (Wired, 3 March 2014) <www.wired.com/2014/03/bitcoin-exchange/#:~:text=Tokyo%2Dbased%20bitcoin%20exchange%20Mt,the%20much%2Dhyped%20digital%20currency> accessed 11 May 2021.

('PSA') in 2017, and created a licensing regime for its cryptocurrency exchanges.[185] According to the new amendment, all cryptocurrency exchanges engaged in purchasing, selling, and exchanging cryptocurrencies and intermediaries thereof, must register with the Japan Financial Services Authority ('FSA').[186] The exchanges must also comply with capital and positive net assets requirements and have an internal auditing system to ensure compliance with the relevant PSA rules.[187] To date, the FSA has granted licenses to 16 cryptocurrency exchanges subjecting them to regulatory oversight.[188]

Like Japan, the licensing regime will help the US regulator supervise the cryptocurrency intermediaries. The regime will establish a market oversight mechanism by ensuring that these intermediaries comply with the relevant US laws, including AML/CFT, anti-manipulation, anti-fraud, consumer disclosure, and prudential (licensing and minimum-capitalisation) requirements.

(iii) A Central Cryptocurrency Trading Platform with Registration Requirement

Under the new cryptocurrency regulation, there should be a central cryptocurrency trading platform with mandatory registration requirements for all cryptocurrency exchanges and spot markets willing to trade cryptocurrencies and issue cryptocurrency-based offerings. A centrally regulated cryptocurrency trading platform with a mandatory registration requirement will infuse the liquidity in the spot market and stabilise the cryptocurrency's price volatility.[189] This will further facilitate the regulators to exercise direct oversight over the cryptocurrency spot markets. A central platform predicated on disclosure requirements and information-sharing will provide investors with transparency and equal access to information, and will safeguard the market from the cryptocurrency world, which is highly asymmetrical, unverified, and sometimes blatantly false.[190]

As to the platform's structure, the new agency could create and maintain a central database where all cryptocurrency-based offerings will be listed and

[185] Masahiko Ishida, Edward Mears, and Ryutaro Takeda, 'Japan Regulatory Update on Virtual Currency Business' (DLA Piper, 29 December 2017) <www.dlapiper.com/en/japan/insights/publications/2017/12/japan-regulatory-update-on-virtual-currency-business/> accessed 23 November 2020.

[186] ibid.

[187] ibid.

[188] These cryptocurrency exchanges are represented by the Japan Virtual Exchange Association—a self-regulatory organization for the Japanese cryptocurrency industry.

[189] Brial Novell, 'Regulation Crypto—A Proposed Framework for United States Cryptocurrency Regulation' (PR Newswire, 20 August 2018) <www.prnewswire.com/news-releases/regulation-crypto--a-proposed-framework-for-united-states-cryptocurrency-regulation-300699235.html> accessed 23 November 2020.

[190] ibid.

conducted, providing the regulator "with direct oversight, capabilities, and transparency for all transactions and parties involved".[191] Further, to prevent false claims on ICOs, crypto-derivatives, and other cryptocurrency-based offerings, here should be a requirement that all advertisements and solicitations be source-verified.[192] The registration and offering process should be based on disclosure requirements, representations, and warranties, as well as the regulator's involvement at the outset. In addition to supporting, facilitating, and listing ICOs and crypto-derivatives, a central cryptocurrency exchange could employ world-class security measures, which would consequently enhance the trading platform's safety and investors' confidence.[193]

V. Conclusion

There are many questions yet to be answered in this rapidly changing cryptocurrency industry. However, hype, volatility, and speculation — these three words best describe the current cryptocurrency space, which is likely to adversely impact the crypto-derivatives investors in the absence of a robust regulatory framework, investor protection, and oversight mechanisms of the spot market.

The UK's and the EU's regulators have recently moved to protect retail investors from the risk and volatility of the crypto-derivatives market, and have extended their efforts to establish a uniform legislative framework for all kinds of cryptocurrency-related assets. Nevertheless, the US offers a somewhat hands-off approach. The present study takes a comparative approach to analyse the regulators' contrasting stance with respect to crypto-derivatives in the UK, EU, and US. The research revolves around an essential question — whether the existing regulatory approach towards crypto-derivatives is adequate to protect the retail investors? It concludes that unlike the UK and the EU, the US measures fail considerably to consider the 'Main Street' investors' vulnerability to this allegedly over-leveraged crypto-derivatives market. It further concludes that the propensity of market manipulation in the cryptocurrency spot market combined with regulatory opacity and fragmentation creates significant hurdles to regulate crypto-derivatives under the existing US legal framework.

These opaque and fragmented regulatory responses to crypto-derivatives demonstrate the dire need for centralised and comprehensive cryptocurrency regulation in the US. The lack of regulation perpetuates the fraudulent and manipulative behaviour in the spot market, that will eventually drive potential investors away from the market. Many crypto-enthusiasts resist the idea of

[191] ibid.
[192] ibid.
[193] ibid.

regulating the industry on the ground that it may impede innovation. However, regulatory clarities and a proper oversight mechanism over the cryptocurrency exchanges and spot markets are necessary frictions, as this will help protect the market integrity, punish abusive and manipulative practices, and restore investor confidence.

To reimagine the crypto-derivatives regulation in the US, the paper envisions a centralised disclosure-based cryptocurrency regulatory regime in establishing an effective oversight mechanism for the US cryptocurrency spot market. This would not increase regulatory certainties and competencies and is also likely to attract a wide range of market participants.

At the 5th Annual Conference on FinTech and Regulation held in February 2021, Robert Ophèle, as Chairman of Autorité des Marchés Financiers, proposed a centralised crypto-regulation on the ground that a single regulator is cheaper and more competent to exercise the centralised expertise in this emerging cryptocurrency market.[194] In the US, Hester Peirce, as an SEC Commissioner, called for regulatory clarities as major corporations like Tesla, BNY Mellon, and Mastercard, started participating in the cryptocurrency market.[195]

Therefore, like the UK and the EU, the US Congress should develop comprehensive federal cryptocurrency legislation to capture the cryptocurrency's novelty and underpinning technology. A comprehensive regulation will serve the public interest, providing a systemic regulatory approach that minimises the risks of cryptocurrency spot markets and averts market failure. Furthermore, the legislation will encapsulate the technology-specific regulation to cryptocurrency, strengthening the US capital and financial market. A uniform federal-level cryptocurrency agency is likely to establish effective oversight and supervisory authority over the cryptocurrency spot markets, that would help curb market manipulation and protect retail investors, and thereby uphold market integrity.

[194] In Robert Ophèle's opinion, a single body could provide a level playing field among all cryptocurrency service providers and is likely to have all the expertise that would provide simplicities and certainties in regulating cryptocurrency. In his view, the ESMA, in the EU, could be a competent authority to oversee cryptocurrency spot markets and any financial instruments where the reference asset is a cryptocurrency. See Helen Partz, 'French Official Wants to Change How Europe Regulates Crypto and Blockchain' (The CoinTelegraph, 09 February 2021) <https://cointelegraph.com/news/french-official-wants-to-change-how-europe-regulates-crypto-and-blockchain> accessed 20 February 2021.

[195] Chris Prentice and Katanga Johnson, 'Clear Crypto Rules Urgently Needed as Major Companies Embrace Asset: SEC Official' (The Reuters, 13 February 2021) <www.reuters.com/article/idUSKBN2AD0ML> accessed 24 February 2021.

Cambridge Law Review (2021) Vol VI, Issue i, 137–181

Are Involuntary Creditors Adequately Protected from the Adverse Impact of the Doctrine of Limited Liability? An Analysis of the Origins of the Doctrine and its Modern Application Through the Prism of Involuntary Creditors' Protection

Mikołaj kudliński*

Abstract

The doctrine of limited liability is considered as one of the most important issues in corporate law. This is because, by limiting shareholders' exposure to risk, limited liability incentivises people to invest in corporate entities and pursue various business endeavours, which in turn stimulates economic growth. However, it is also often argued that the doctrine of limited liability is controversial, as it allows companies to easily externalise their commercial risks, which exposes the vulnerable group of involuntary creditors to significant losses. This problem is particularly evident in the context of corporate groups, where parent companies use the corporate form to insulate themselves from liability for the acts of their subsidiaries. This paper discusses the origins of the limited liability doctrine through the prism of its development in the United Kingdom, and finds that that the interests of involuntary creditors were not given adequate consideration at the time of its

* Christ Church, University of Oxford. LLB Hons (Aberdeen), LLM (Edinburgh). I am grateful to Adam Strukowski and Izabella Kuna for their help in the revision of this article. I am also indebted to the anonymous editors of Cambridge Law Review for their valuable comments on the earlier draft. Any errors remain my own. mikolaj.kudlinski@chch.ox.ac.uk.

inception. Arguably, this doctrine was never supposed to be applied in relation to this group of creditors at all. Subsequently, this paper discusses the current protection mechanism available to involuntary creditors in the United Kingdom and finds that, for various reasons, these mechanisms are not effective. This article concludes by discussing alternative approaches to limited liability and noting that the control-based presumption of parent liability would strike a fair balance between the interests of the various actors involved in the company's activity, and would provide involuntary creditors with a greater degree of protection.

Keywords: company law, limited liability, creditors, involuntary creditors, alternatives to limited liability

I. Introduction

Companies form an inextricable part of modern society. Today, corporate entities are encountered virtually everywhere; companies produce and distribute countless products, provide various services, run all types of transport, supply weapons, engage in politics, and have the ability to influence global financial markets.[1] Moreover, companies are seen as the main drivers of the globalisation process.[2] For the above reasons, corporations are "among the most powerful institutions of our time".[3]

Companies underpin the capitalist economies upon which modern societies are predicated.[4] In fact, in 2019, companies represented 72.5% of total businesses in the United Kingdom.[5] Likewise, at the end of June 2020, there were four million five hundred thousand three hundred ninety-two corporate entities on the total register of companies; four million one hundred thousand three hundred twenty-

1 Sarah Worthington, *Sealy and Worthington's Text, Cases, and Materials in Company Law* (Oxford University Press 2016) 1.

2 Stuart Kirsch, *Mining Capitalism: The Relationship Between Corporations and Their Critics* (University of California Press 2014) 1.

3 ibid. On this account, Noam Chomsky described the most powerful corporations as today's "masters of mankind". See Noam Chomsky, *Masters of Mankind: Essays and Lectures, 1969-2013* (Haymarket Books 2014).

4 Peter A Hall and David Soskice, 'An Introduction to Varieties of Capitalism' in Peter A. Hall and David Soskice (eds), *Varieties of Capitalism - The Institutional Foundations of Comparative Advantage* (Oxford University Press 2004) 6.

5 Office for National Statistics, 'UK Business: Activity, Size and Location 2019' (2 October 2019) 4 <www.ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/bulletins/ukbusinessactivitysizeandlocation/2019> accessed 25 May 2020.

three were actively trading.[6] The number of companies on the total register is growing at a steady rate and has increased since 1979 by over three million four hundred thousand.[7] During the same period, the number of companies on the effective register has grown by over three million two hundred thousand.[8] This data evidences that companies are the main vehicles through which business is carried out in the United Kingdom.[9]

Upon incorporation, a company becomes a legal person distinct from its shareholders.[10] According to Armour et al., being a legal fiction, the company can: (a) enter into contracts; (b) have rights in property; (c) sue and be sued in its own name and; (d) delegate authority to agents.[11] Thus, as a consequence of incorporation, the company has its own rights and is capable of undertaking its own obligations.[12] The doctrine of separate legal personality is therefore "fundamental"[13] to the conceptual understanding of the structure of corporate law. In addition, it is of crucial significance from the functional perspective, as the separate legal personality of the company makes it possible to distinguish the assets owned by the company's members from the assets owned by the company itself.[14]

6 Companies House, 'Official Statistics: Incorporated companies in the UK April to June 2020' (30 July 2020) <www.gov.uk/government/publications/incorporated-companies-in-the-uk-april-to-june-2020/incorporated-companies-in-the-uk-april-to-june-2020> accessed 4 August 2020.

7 Companies House, 'Official Statistics: Companies register activities: 2018 to 2019' (1 August 2019) Section 3. <www.gov.uk/government/publications/companies-register-activities-statistical-release-2018-to-2019/companies-register-activities-2018-to-2019#overseas> accessed 25 May 2020.

8 ibid.

9 Similarly, in Australia, between 2018 and 2019, companies represented 37.9% of all businesses, the largest of any type of legal organisation. See Australian Bureau of Statistics, '8165.0 - Counts of Australian Businesses, including Entries and Exits, June 2015 to June 2019' (20 February 2020) <www.abs.gov.au/ausstats/abs@.nsf/mf/8165.0> accessed 25 May 2020.

10 *Salomon v Salomon* [1897] AC 22 HL at [31] per Lord Halsbury, at [51] per Lord Macnaghten. See also Brenda Hannigan, *Company Law* (5th edn., Oxford University Press 2018) 42.

11 John Armour, Henry Hansmann, Reinier Kraakman and Mariana Pargendler, 'What is Corporate Law?' in Kraakman et al. (eds), *Anatomy of Corporate Law: A Comparative and Functional Approach* (Oxford University Press 2017) 8.

12 Phillip I Blumberg, 'Limited Liability and Corporate Groups' (1986) 11 Journal of Corporate Law 573, 577.

13 Paul Davies, *Introduction to Company Law* (Oxford University Press 2002) 9.

14 ibid 11.

This, in turn, facilitates the operation of the doctrine of limited liability, which constitutes another elemental principle of corporate law.[15]

The doctrine of limited liability is present in almost all developed legal systems in the world.[16] It presupposes that the liability of the company's members is restricted to the amount they have agreed to pay for the company's shares.[17] Consequently, shareholders will not be held personally liable for the debts of the company.[18] Thus, the existence of limited liability encourages investors to pursue business endeavours, which otherwise could be regarded as being too risky.[19] For that reason, limited liability is widely seen as a mechanism which incentivises entrepreneurship and stimulates economic development.[20]

Due to its pivotal role in encouraging business activity, the doctrine of limited liability is of crucial economic importance and has been excessively praised in this context. In fact, limited liability is often regarded as "the most important characteristic of the modern corporation",[21] "the hallmark of corporate status",[22] or "an unsung hero"[23] of free market economies. Moreover, in their appraisal of limited liability, some commentators have asserted that, in the historical development of the corporation, "no single attribute had been more significant than limited liability",[24] or that limited liability is "the most effective legal invention

15 Frank H Easterbrook and Daniel R Fischel, 'Limited Liability and the Corporation' (1985) 52(1) University of Chicago Law Review 89, 89.

16 Sung Bae Kim, 'A Comparison of the Doctrine of Piercing the Corporate Veil in the United States and in South Korea' (1995) 3 Tulsa Journal of Comparative and International Law 73, 73.

17 Armour, Hansmann, Kraakman and Pargendler (n 11) 6-10; David Kershaw, *Company Law in Context: Text and Materials* (2nd edn., Oxford University Press 2012) 20; Paul Davies and Sarah Worthington, *Gower's Principles of Modern Company Law* (10th edn., Sweet & Maxwell/Thomson Reuters 2016) 191; Easterbrook and Fischel (n 15) 89-90.

18 Davies (n 13) 60; Stefan H C Lo, 'Liability of Directors as Joint Tortfeasors' (2009) 2 Journal of Business Law 109, 119.

19 Stephen Griffin, 'Limited Liability: A Necessary Revolution' (2004) 25(4) The Company Lawyer 99, 99.

20 ibid; Andrew Hicks, 'Corporate Form: Questioning the Unsung Hero' (1997) Journal of Business Law 306, 306-307; Judith Freedman, 'Limited Liability: Large Company Theory and Small Firms' (2000) 63(3) The Modern Law Review 317, 317.

21 Stephen M. Bainbridge and M. Todd Henderson, *Limited Liability: A Legal and Economic Analysis* (Edward Elgar Publishing 2016) 19.

22 Christopher W Peterson, 'Piercing the Corporate Veil by Tort Creditors' (2017) 13 Journal of Business and Technology Law 63, 63.

23 Institute of Directors, Deregulation for Small Private Companies (IOD, 1986) quoted in Judith Freedman, 'Small Businesses and the Corporate Form: Burden or Privilege?' (1994) 57(4) Modern Law Review 555, 564.

24 Warner Fuller, 'The Incorporated Individual: A Study of the One-Man Company' (1938) 51 Harvard Law Review 1373, 1376.

of the nineteenth century".[25] Others have described the limited liability company as "the greatest single discovery of modern times",[26] which is even more important than steam or electricity. Accordingly, the nameless creator of limited liability deserves a "place of honour" among the pioneers of the Industrial Revolution, such as Watt or Stephenson.[27]

Certain scholars consider limited liability, however, as one of "the most controversial issues in corporate law".[28] In fact, it is argued that since the introduction of general limited liability into the UK law in the nineteenth century, creditors dealing with companies have been exposed to excessive risks.[29] Indeed, business activity has intrinsic costs, which upon incorporation are externalised onto the company's creditors.[30]

Within the wider group of corporate creditors, one can distinguish voluntary and involuntary creditors.[31] Voluntary creditors are able to determine the creditworthiness of a particular company and gauge the risks that could arise from their dealings with such a company.[32] Consequently, they are able to bargain with the corporation and protect themselves from the aforesaid risks through,

[25] President Charles William Eliot of Harvard University quoted in William W. Cook, 'Watered Stock Commissions Blue Sky Laws Stock Without Par Value' (1921) 19(6) Michigan Law Review 583, 583.

[26] President Nicholas Murray Butler of Columbia University quoted in William Meade Fletcher, Cyclopedia of the Law of Private Corporations (Callaghan and Company 1917) 43, 21.

[27] The Economist, 18 December 1926 quoted in Bishop Carleton Hunt, *The Development of the Business Corporation in England, 1800-1867* (Harvard University Press 1936) 116.

[28] Larry E Ribstein, 'Limited Liability and Theories of the Corporation' (1991) 50 Maryland Law Review 80, 81; Colin Mackie, 'From Privilege to Right: Themes in the Emergence of Limited Liability' (2011) 4 Juridical Review 293, 294.

[29] Bob Tricker, 'Re-Inventing the Limited Liability Company' (2011) 19(4) Corporate Governance: An International Review 384, 385-386.

[30] David Millon, 'Piercing the Corporate Veil, Financial Responsibility, and the Limits of Limited Liability' (2006) 56 Emory Law Journal 1305, 1355.

[31] Andrew Muscat, *The Liability of the Holding Company for the Debts of its Insolvent Subsidiaries* (Routledge 2016) at 4.5; Peter French, 'Parent Corporation Liability: An Evaluation of the Corporate Veil Piercing Doctrine and its Application to the Toxic Tort Arena' (1992) 5(2) Tulane Environmental Law Journal 605, 607.

[32] Henry Hansmann and Reinier Kraakman, 'Toward Unlimited Shareholder Liability for Corporate Torts' (1991) 100(7) Yale Law Journal 1879, 1919-1920.

for instance, specific contractual arrangements, securing guarantees, or charging higher rates for the credit.[33]

Involuntary creditors, on the other hand, cannot 'choose their tortfeasor' and have no means of allocating the risk of losses or injury.[34] They are therefore poor risk-bearers.[35] Moreover, involuntary creditors can only seek compensation *ex post*.[36] Because the doctrine of limited liability shields the assets of the company's owners, the wrongdoing company often may lack sufficient funds to pay out damages to the involuntary creditors affected by its actions.[37] In such instances, the creditors may be left with no compensation at all.[38] The weak position of involuntary creditors is further exacerbated by the fact that limited liability enables parent companies within a corporate group to avoid responsibility for the harm caused by their subsidiaries.[39] Namely, even though the parent company may effectively control the subsidiary, they are separate persons in the eyes of the law.[40] The parent company will therefore not be held liable for the debts of its subsidiary.[41] As a result, the subsidiary may not have enough assets to cover its liabilities.[42] For instance, in *Adams v Cape Industries plc*,[43] the parent company was able to escape liability for the debts of its insolvent subsidiary and Mr Adams, who contracted asbestosis as a result of his employment with the subsidiary, was left with almost nothing.

Another dire consequence of limited liability for involuntary creditors, and society at large, is the fact that the doctrine incentivises opportunism and corporate recklessness.[44] Namely, it is widely contended that limited liability encourages

[33] See Phillip Lipton, 'The Mythology of Salomon's Case and the Law Dealing With the Tort Liabilities of Corporate Groups: An Historical Perspective' (2014) 40 Monash University Law Review 452, 481; Lo, 'Liability of Directors as Joint Tortfeasors' (n 18) 121.

[34] Lipton (n 33) 481.

[35] Robert B Thompson, 'Piercing the Corporate Veil: An Empirical Study' (1991) 76 Cornell Law Review 1036, 1070-1073; French (n 31) 608-609.

[36] Millon (n 30) 1355.

[37] Lo, 'Liability of Directors as Joint Tortfeasors' (n 18) 121.

[38] ibid.

[39] Blumberg (n 12) 575; Peter Muchlinski, 'Limited Liability and Multinational Enterprises: a Case for Reform?' (2010) 34(5) Cambridge Journal of Economics 915, 915-916; Lipton (n 33) 480-481; Martin Petrin and Barnali Choudhury, 'Group Company Liability' (2018) 19(4) European Business Organization Law Review 771, 773-774.

[40] Andreas Rühmkorf, *Corporate Social Responsibility, Private Law and Global Supply Chains* (Edward Elgar Publishing 2015) 172.

[41] ibid.

[42] ibid.

[43] *Adams v Cape Industries plc* [1990] Ch. 433.

[44] Hansmann and Kraakman (n 32) 1920; Ribstein (n 28) 81.

companies to engage in hazardous behaviour,[45] as the potential liability of their owners is considerably restricted.[46] For example, following the Bhopal disaster in December 1984, which claimed the lives of over twenty-two thousand people, in 2012, a US district court held that Union Carbide Corporation had no liability related to the plant site owned by its Indian subsidiary, and thus could not be held liable for any pollution-related damage.[47] This was, *inter alia*, because the parent company and the subsidiary were separate persons in the eyes of the law.

For the above reasons, it is argued that the application of the doctrine of limited liability has been extended beyond its original purpose.[48] It is no longer used as a mechanism that stimulates business activity, but is rather used as a vehicle through which "irresponsibility is institutionalised".[49] Today, limited liability allows companies to benefit from the diversification of risks, which in turn exposes involuntary creditors to an excessive danger of loss or harm.[50] In this regard, it is worth noting that the existence of insurance cannot address this problem on its own. Namely, it can hardly be argued that involuntary creditors could predict their injury, and thus could insure themselves, before they were injured. Likewise, from a practical point of view, it is virtually impossible to insure every potential

[45] For instance, in a type of behaviour that is dangerous to the environment, see Nick Grant, 'Mandating Corporate Environmental Responsibility by Creating a New Directors' Duty' (2015) 17(4) Environmental Law Review 252, 252-254.

[46] Davies and Worthington (n 17) 194; Bainbridge and Henderson (n 21) 49-51; Carsten Gerner-Beuerle and Michael Anderson Schillig, *Comparative Company Law* (Oxford University Press 2019) 46; David Campbell and Stephen Griffin, 'Enron and the End of Corporate Governance?' in Sorcha MacLeod (ed), *Global Governance and the Quest for Justice: Volume II Corporate Governance* (Hart Publishing 2006) 48; Easterbrook and Fischel (n 15) 103-104; David W. Leebron, 'Limited Liability, Tort Victims, and Creditors' (1991) 91(7) Columbia Law Review 1565, 1565; Andrew Price, 'Tort Creditor Superpriority and Other Proposed Solutions to Corporate Limited Liability and the Problem of Externalities' (1995) 2 George Mason Law Review 439, 441-442; Muchlinski, 'Limited Liability and Multinational Enterprises: a Case for Reform?' (n 39) 915-916; Lipton (n 33) 480-481.

[47] *Janki Bay Sahu and others v Union Carbide Corporation and Warren Anderson* (2012) No. 04 Civ. 8825 (JFK). On the legal aspects of the Bhopal disaster, see Jamie Cassels, 'The Uncertain Promise of Law: Lessons from Bhopal' (1991) 29(1) Osgoode Hall Law Journal 1.

[48] Blumberg (n 12) 575.

[49] Paddy Ireland, 'Limited Liability, Shareholder Rights and the Problem of Corporate Irresponsibility' (2010) 34(5) Cambridge Journal of Economics 837, 838. See also Tricker (n 29) 386.

[50] Charlotte Villiers, 'Corporate Law, Corporate Power and Corporate Social Responsibility' in Nina Boeger, Rachel Murray and Charlotte Villiers (eds), *Perspectives on Corporate Social Responsibility* (Edward Elgar Publishing 2008) 95.

involuntary creditor against every potential risk of injury. Moreover, in certain large cases, a company's liability may considerably exceed its insurance coverage.[51]

Given the aforementioned factors, this article considers the question whether involuntary creditors are appropriately protected by the law of the United Kingdom. This will be done through the prism of the origins of the doctrine of limited liability and its modern application. Part II of this paper will analyse the manner in which the doctrine of limited liability has developed in the UK, and whether the interests of involuntary creditors were given adequate consideration at the time of its inception. Part III will outline the economic rationale behind limited liability and will examine the protection mechanisms available to involuntary creditors today, such as piercing the corporate veil, bypassing limited liability under tort law, and s.172(1) of the Companies Act 2006. It will be concluded that the current protection mechanisms are not effective, and that involuntary creditors are continuously exposed to excessive risks. On this account, Part IV will evaluate the alternatives to limited liability from the perspective of involuntary creditors, such as pro rata liability of shareholders, giving preference to involuntary creditors on insolvency, and the control-based liability system coupled with the control-based presumption of parent liability. Arguably, involuntary creditors would be afforded a greater degree of protection under the last of these approaches, which, given the slow transition process of the UK economy from a pure profit-orientated system towards a more stakeholder-inclusive one, could potentially be introduced in the future.

II. The Historical Development of the Doctrine of Limited Liability in the United Kingdom: An Analysis Through the Prism of Involuntary Creditors' Protection

Jean du Plessis begins his analysis of the history of UK company law by stating that "[i]t cannot be disputed that corporate law cannot be understood without a proper knowledge of the historical context in which it developed".[52] In this statement, du Plessis posits that an understanding of the historical

[51] Cassels (n 47) 9.

[52] Jean du Plessis, 'Corporate Law and Corporate Governance Lessons From the Past: Ebbs and Flows, But Far From the "End of History": Part 1' (2009) 30(2) Company Lawyer 43, 45.

development of company law is necessary for the proper comprehension of its modern framework.[53] Later, du Plessis adds, quoting William Ashley,[54] that

> "[i]n corporate law history [...] it has been observed that 'in every stage of social evolution there are particular needs which have to be met, and particular tendencies in human character which call either for repression or stimulus'".[55]

In the above extract, du Plessis argues that company law has developed in certain patterns, which reflected the specific needs of society at a given point in time.[56] In sum, in the foregoing passages, du Plessis suggests that modern company law cannot be properly understood without a prior understanding of the "particular needs"[57] of society that corporate law has attempted to remedy throughout its history. This is particularly true when one embarks upon an analysis of the development of the doctrine of limited liability in the United Kingdom, which was highly influenced by the peculiar social, economic, and political climate of contemporary Britain.[58] Accordingly, the following paragraphs will consider, through the lens of involuntary creditors' protection, the manner in which the

[53] John Armour agrees that registered companies cannot be properly conceptualised under English law without the prior understanding of their historical development: John Armour, 'Companies and Other Associations' in Andrew Burrows (ed), *English Private Law* (Oxford University Press 2013) at 3-44.

[54] William J Ashley, *An Introduction to English Economic History and Theory* (3rd edn, Vol.1, Longmans Green & Co. 1894) 167-168.

[55] Jean du Plessis, 'Corporate Law and Corporate Governance Lessons From the Past: Ebbs and Flows, But Far From the "End of History": Part 2' (2009) 30(3) Company Lawyer 72, 73.

[56] Jonathan C Hardman, 'Resolving Agency Costs in United Kingdom Private Companies' (Phd thesis, University of Glasgow 2020) 72.

[57] ibid.

[58] See Mackie (n 28) 294-295.

concept of limited liability has evolved in the UK and the purposes for which it was introduced.

A. The origins of the doctrine of limited liability in the UK

Arguably, the principle of limited liability can be traced back to Roman times.[59] In this context, legal historians often invoke the concept of a *peculium*.[60] Moreover, Roman law recognised certain non-human entities, such as *universitas personarum*, which were treated by law like real persons; such entities had legal capacity and could undertake their own duties.[61] In this regard, Ulpian's well-known maxim stated that where something was owed by the 'corporation', it was not owed by its members, but by the 'corporation' itself.[62] This maxim can therefore be considered as an early promulgation of the doctrines of separate legal personality and limited liability.

In medieval and early modern England, corporations were "direct outgrowths of the state" created under Royal Charters.[63] The early corporations, such as guilds, existed almost exclusively for the benefit of the public, and not for the private benefit of their members.[64] Due to their intrinsic public purpose and their public functions, the early corporations did not engage in risky profit-generating endeavours and rarely incurred debt.[65] For that reason, the early corporate forms

59 Robert W Hillman, 'Limited Liability in Historical Perspective' (1997) 54 Washington and Lee Law Review 615, 616-619.

60 This mechanism allowed the head of a Roman family ("Paterfamilias" in Latin) to entrust his slave, or another member of his family, with an amount of capital in order for the grantee to carry out trading on his behalf. Crucially, the assets remained the property of the paterfamilias, and he was only liable for the debts incurred by the grantee in the course of business to the extent of the peculium. The peculium was therefore an excellent limited liability mechanism, which facilitated commerce: see Henry Sumner Maine, *Ancient Law and the Origins of Human Society* (5th edn., John Murray, London 1888) 129; Bainbridge and Henderson (n 21) 22; David V. Snyder, 'The Case of Natural Obligations' (1995) 56 Louisiana Law Review 423, 429; Hillman (n 59) 616-619.

61 Laura Macgregor, 'Partnerships and Legal Personality: Cautionary Tales from Scotland' (2020) 20(1) Journal of Corporate Law Studies 237, 248-249. See also Leonardo Davoudi, Christopher McKenna and Rowena Olegario, 'The Historical Role of the Corporation in Society' (2018) 6.s1 Journal of the British Academy 17, 22-23.

62 "*Si quid universitate debetur singuli non debetur; nec quod debet universitas singuli debent*": The Digest of Justinian 3.4.7.1. (Ulpianus).

63 Bainbridge and Henderson (n 21) 27.

64 Colin Arthur Cooke, *Corporation, Trust & Company: An Essay in Legal History* (Manchester University Press 1950) 51; Hardman (n 56) 73.

65 Phillip Lipton, 'The Introduction of Limited Liability into the English and Australian Colonial Companies Acts: Inevitable Progression or Chaotic History' (2017) 41 Melbourne University Law Review 1278, 1286.

had no need to externalise risks, and limited liability was not of major significance to them.[66]

Subsequently, together with the rapid geographical expansion of the 16th and 17th centuries, commercial joint-stock companies emerged with an aim to accumulate capital from investors in order to fund business activity overseas.[67] Notably, these associations were established with the purpose to maximise the private financial gains of their members, as the development of foreign trade significantly increased the potential for wealth creation.[68] Business activity abroad, however, was associated not only with substantial gains but also with considerable political and commercial risks.[69] For that reason, the problem of investors' liability became an important issue at the time.[70] One of the possibilities available to investors in order to minimise the risk of future liabilities was incorporation, which could be obtained either under a Royal Charter or an Act of Parliament.[71]

Upon incorporation, a joint-stock company became an artificial person and the charter conferred numerous benefits, such as the monopoly of trade, upon the corporation's members.[72] Incorporation was therefore regarded as an important privilege, which was "jealously guarded".[73] In fact, the process of obtaining a charter was expensive and notoriously difficult.[74] Furthermore, after the introduction of the Bubble Act 1720, incorporating a company became even harder and unincorporated joint-stock companies were declared illegal.[75] As a result, the majority of contemporary joint-stock companies traded without incorporation through the use of trusts, and their functioning was regulated by the

66 ibid 1286-1287.

67 Igho Lordson Dabor, 'Limited Liability: A Pathway for Corporate Recklessness?' (Phd thesis, University of Wolverhampton 2016) 17.

68 ibid.

69 Vyuptakesh Sharan, *International Business: Concepts, Environment And Strategy* (Pearson Education India 2008) 9. See also Markéta Kadlecová, 'England and the Promotion of Trade in 16th and 17th Centuries' (2014) 2(4) West Bohemian Historical Review 13, 13-28.

70 Lipton (n 65) 1286-87.

71 Laurence C B Gower, 'The English Private Company' (1953) 18 Law and Contemporary Problems 535, 535.

72 Armour (n 53) 3-37, 3-45.

73 Frederick G Kempin Jr, 'Limited Liability in Historical Perspective' (1960) 4(1) American Business Law Association Bulletin 11, 13.

74 Hardman (n 56) 73-74; Adrian Henriques, Corporate Impact: *Measuring and Managing Your Social Footprint* (Earthscan 2010) 14.

75 In fact, the purpose for which the Bubble Act 1720 was introduced is debatable: Dabor (n 67) 50; Hardman (n 56) 74-76; Kirstin Olsen, *Daily Life in 18th-century England* (2nd edn., ABC-CLIO 2017) 107. See also Ron Harris, 'The Bubble Act: Its Passage and its Effects on Business Organization' (1994) 54(3) Journal of Economic History 610, 610-627.

relevant deeds of settlement.[76] Such companies, however, were *de facto* and *de jure* unlimited.[77]

In regard to incorporated companies, the treatment of limited liability at that time was unclear.[78] Various scholars assert that, by the end of the seventeenth century, limited liability was seen as a common benefit of incorporation.[79] For instance, in 1662, the Act Declaratory Concerning Bankrupts provided the shareholders of the East India Company, the Royal African Company, and the Royal Fisheries Company with a form of limited liability from losses incurred by these companies in the course of trading.[80] Likewise, in *Edmunds v Brown and Tillard*, the King's Bench held that members of a chartered corporation could not be held liable, in their personal capacities, for debts of the dissolved corporation.[81]

Conversely, it is also argued that limited liability was not "a necessary incident of incorporation" at the time,[82] and thus the concept was relatively unimportant.[83] Indeed, in defining the characteristics of the corporation, the contemporary legal commentary did not refer to limited liability.[84] For example, Blackstone wrote that an incorporated body had perpetual succession, it could sue and be sued, it could have rights in property, and it had a common seal.[85] Blackstone did not, however, make a reference to limited liability in his work. According to Blumberg, neither Sir Edward Coke,[86] writing before Blackstone, nor Stewart Kyd,[87] writing after him,

[76] Armour (n 53) 3-44-3-45; John D Turner, 'The Development of English Company Law Before 1900' in Harwell Wells (ed), *Research Handbook on the History of Corporate and Company Law* (Edward Elgar Publishing 2018) 128-129; Hardman (n 56) 76.

[77] Turner (n 76) 129.

[78] Armour (n 53) 3-38.

[79] William Searle Holdsworth, *A History of English Law* (5th edn., Sweet & Maxwell 1973) 484; Laurence Cecil Bartlett Gower et al., *Principles of Modern Company Law* (4th edn., Stevens and Sons 1979) 26; Lipton (n 65) 1287. See also Samuel Williston, 'History of the Law of Business Corporation Before 1800' Part 2 (1888) 2.3 Harvard Law Review 149, 161-162.

[80] William R Scott, *The Constitution and Finance of English, Scottish and Irish-Joint Stock Corporations to 1720* (Cambridge University Press 1910–12) 1, 270; Lipton (n 65) 1287.

[81] *Edmunds v Brown and Tillard* (1667) 83 E.R. 385.

[82] Robert A Kessler, 'With Limited Liability for All: Why Not a Partnership Corporation?' (1967) 36(2) Fordham Law Review 235, 241.

[83] Armand Budington DuBois, *The English Business Company after the Bubble Act, 1720–1800* (Commonwealth Fund 1938) 93-94; Merrick E Dodd, 'The Evolution of Limited Liability in American Industry: Massachusetts' (1948) 61 Harvard Law Review 1351, 1351; Blumberg (n 12) 579.

[84] Blumberg (n 12) 579-580.

[85] Sir William Blackstone, *Commentaries on the Laws of England*, Book 1 (Oxford Clarendon Press 1753) 464-465.

[86] Edward Coke, *First Part of the Institute of the Laws of England or A Commentary Upon Littleton* (1628) 6, 412.

[87] Stewart Kyd, *Corporations* (1793) 103.

listed limited liability as an essential characteristic of the corporation.[88] Moreover, in the *Sutton's Hospital* case, the King's Bench did not include limited liability in its list of the crucial features of a corporate body.[89] In addition, despite the general view that the debts of a corporation were not the debts of its members, in *Dr. Salmon v The Hamborough Company,* the House of Lords held that when the charter so specifies, the corporation could make a "leviation" upon its members in order to satisfy its liabilities towards the creditors.[90] Therefore, the above observations support the view that the instances where limited liability applied at the time were in fact rare,[91] and the actual meaning of this doctrine was unclear.[92]

It follows from the foregoing that the treatment of limited liability in the 17th and 18th centuries was ambiguous.[93] Arguably, the existence of limited liability was dependent upon the wording of the relevant charter, and whether this privilege was expressly granted therein.[94] The Crown, however, took an inconsistent approach towards inserting limited liability clauses in the incorporation charters; some charters contained an express clause limiting direct shareholder liability, and others did not.[95] Nevertheless, according to Harris, the benefit of limited liability associated with incorporation was slowly becoming a crucial motive for obtaining a charter towards the end of the eighteenth century.[96] In fact, when the Warmley Company applied for incorporation in 1768,[97] its representative stated in the

[88] Blumberg (n 12) 579-580.

[89] *Sutton's Hospital* (1612) 10 Coke Reports 23a. See also Kessler (n 82) 240.

[90] *Dr. Salmon v The Hamborough Company* (1671) 22 E.R. 763 at [764]. See Armour (n 53) 3-38.

[91] DuBois (n 83) 94.

[92] Turner (n 76) 131.

[93] Blumberg (n 12) 578-580. See also Oscar Handlin and Mary F Handlin, 'Origins of the American Business Corporation' (1945) 5(1) Journal of Economic History 1, 9-10.

[94] Ron Harris, *Industrializing English Law: Entrepreneurship and Business Organization, 1720-1844* (Cambridge University Press 2000) 128-129; Armour (n 53) 3-38; Turner (n 76) 124.

[95] Blumberg (n 12) 580.

[96] Harris (n 94) 129-131.

[97] The Warmley Company had a leading position in the copper industry in the first half of the eighteenth century. See Barrie Charles Blake-Coleman, *Copper Wire and Electrical Conductors: The Shaping of a Technology* (CRC Press 1992) 101.

petition that limited liability was of substantial importance to the company, and was thus one of the main rationales for incorporation.[98]

Among the main reasons for which shareholders sought limited liability was protection from personal bankruptcy.[99] Therefore, at that time, limited liability was already used as a mechanism through which shareholders externalised the risks of doing business; when the relevant charter provided for limited liability, the commercial risks were shifted, with the approval of the state, onto the corporation's contractual creditors. Notably, limited liability became an even more important consideration in the nineteenth century together with the rapid pace of industrialisation.

B. The nineteenth century debate: were the interests of involuntary creditors adequately considered?

At the beginning of the nineteenth century, business was primarily carried out by partnerships and individuals trading with unlimited liability.[100] At that time, limited liability was still a "rare and dubious privilege",[101] which was available only to certain corporations at the discretion of the state. Due to the lack of the possibilities whereby commercial risks could be reduced, business ventures were considered unsafe, and thus not many people were willing to put their capital at stake.[102] That being said, in the early nineteenth century, an active market for shares emerged for the first time since the Bubble Act 1720.[103] As a result, shares in joint-stock companies became liquid assets that could easily be realised on the market.[104] This new economic reality, stimulated by industrial development, required the law to evolve in a manner that would facilitate enterprise and investment.

The first major change came with the Joint Stock Companies Registration and Regulation Act 1844, which stated that a joint-stock company could become incorporated upon registration, provided that it had satisfied a number of

98 DuBois (n 83) 95; Phillip L Cottrell, *Industrial Finance, 1830-1914: The Finance and Organization of English Manufacturing Industry* (Routledge 2013) 9. See also Ben Pettet, 'Limited Liability—A Principle for the 21st Century?' (1995) 48(2) Current Legal Problems 125, 125-159. Unfortunately for the members of the Warmley Company, the petition was rejected. This was a calamitous event for the company, as it was subsequently dissolved in 1769. See Blake-Coleman (n 97) 102.

99 Lipton (n 65) 1287.

100 ibid 289.

101 Marie-Laure Djelic, 'When Limited Liability was (Still) an Issue: Mobilization and Politics of Signification in 19th-Century England' (2013) 34(5-6) Organization Studies 595, 599.

102 Harris (n 94) 143; Lipton (n 65) 1289.

103 Armour (n 53) 3-47. See also Harris (n 94) 218–223.

104 Armour (n 53) 3-47.

requirements.[105] After the statute was enacted, incorporating a company became much easier and the number of corporate entities in the early nineteenth century started growing gradually.[106] Therefore, there is some merit in the words of John Morley, who wrote that the introduction of the act "marked the rising dawn of an economic era".[107]

Nevertheless, the new act did not provide for general limited liability,[108] and thus shareholders could still be held responsible for debts of the company.[109] Notably, the principle of limited liability was discussed by the Gladstone Committee in the report that preceded the 1844 Act.[110] The main purpose of the report was to "inquire into the state of the laws respecting joint-stock companies with a view to foster public security",[111] as it was contended that the contemporary incorporation regime could be easily abused by fraudsters. In regard to the doctrine of limited liability, the proponents of its general introduction asserted that this principle would "induce persons of respectability"[112] to invest their money in companies, and thus capital would circulate more easily on the market. Moreover, it was argued that limited liability would allow for a better diversification of risks associated with commercial activities.[113] Therefore, the advocates of general limited liability recognised its extensive economic benefits and argued that such benefits could

[105] Joint Stock Companies Registration and Regulation Act 1844, s. 1. See also Mackie (n 28) 295.

[106] Harris (n 94) 272.

[107] John Morley, *The Life of Gladstone* (Macmillan Company 1903) 1, 247.

[108] Alan Dignam and John Lowry, *Company Law* (8th edn., Oxford University Press 2014) 16; Lorraine Talbot, *Critical Company Law* (2nd edn., Routledge 2015) 30; Rob McQueen, *A Social History of Company Law: Great Britain and the Australian Colonies 1854–1920* (Routledge 2016) 45; Armour (n 53) 3-22.

[109] For a good example of how limited liability was treated by the judiciary following the introduction of the Joint Stock Companies Registration and Regulation Act 1844, see *In the Matter of The Sea Fire and Life Assurance Company of The Joint Stock Companies Winding-up Acts, 1848 and 1849* (*Greenwood's Case*) (1854) 43 E.R. 180, where Lord Cranworth said at [188] that "(...) it is clear that the liability to creditors is not materially affected, and the Legislature has not only not exempted the shareholders from their ordinary obligations as partners, but has expressly enacted that they shall remain liable, subject only to the limitation as to three years in a particular case, which is not now in question". See also Lord Turner's comments at [191]-[193].

[110] Parliamentary Committee to Inquire Into the State of the Laws Respecting Joint Stock Companies, with a View to the Greater Security of the Public HC 119 (1841-43) 51.

[111] ibid. See also McQueen (n 108) 45.

[112] Thomas H Bothamley cited in Matthew Baillie Begbie, *Partnership "en Commandite" or Partnership with Limited Liabilities (according to the Commercial Practice of the Continent of Europe and the United States of America) for the Employment of Capital, the Circulation of Wages and the Revival of Our Home and Colonial Trade* (Effingham Wilson 1848) 239-240.

[113] John Duncan cited in Begbie (n 112) 240-241.

be of great advantage to the general public.[114] The opponents of limited liability contended, however, that there was "no immediate occasion"[115] for the change in the existing law, or that such a change was not "expedient".[116] Furthermore, an argument was put forward that the introduction of limited liability would incentivise speculation, and thus the doctrine would constitute a threat to public security.[117] What is worth mentioning here is that the Committee did not consider the risks that limited liability could pose to involuntary creditors. Instead, the discussion revolved around the idea that limited liability would create a considerable threat to the general public due to its potential to incentivise speculation and fraud.

Ultimately, the 1844 Act did not introduce general limited liability and it subsequently turned out to be a failure, as it proved to be economically inefficient and prone to abuse.[118] As Pleydell-Bouverie commented, the 1844 Act "deterred prudent men of capital"[119] from investing, and thus it hampered the economic development of the country.

In the early 1850s, limited liability attracted the attention of the Parliament due to the work of the Christian Socialist movement, whose members sought to improve the living conditions of the working class and to reduce class tensions within the society by, *inter alia*, allowing employees to own shares in their employer's business through the introduction of general limited liability.[120] As a result of their efforts, limited liability became a subject of heated debate in the United Kingdom.[121]

In this debate, opponents of the introduction of general limited liability echoed the argument that general limited responsibility of shareholders would increase the risk of excessive speculation, and would consequently cause a floodgate of bankruptcies.[122] Likewise, some members of the contemporary

[114] ibid. See also McQueen (n 108) 45.

[115] George Larpent cited in Begbie (n 112) 238.

[116] Kirkman Finlay cited in Begbie (n 112) 238.

[117] John Hodgkin cited in Begbie (n 112) 238.

[118] Dabor (n 67) 87-89.

[119] HC Deb 29 June 1855, Vol.139, cols 321-322. This argument was echoed by Richard Malins, HC Deb 29 June 1855, Vol.139, col. 340.

[120] Djelic (n 101) 602-606; Lipton (n 65) 1290. For a comprehensive discussion of the influence of the Christian Socialist movement on the development of the doctrine of limited liability in the UK see Julia Chaplin, 'The Origins of the 1855/6 Introduction of General Limited Liability in England' (Phd thesis, University of East Anglia 2016) 171-251.

[121] Hunt (n 27) 118; William Cornish et al., *Law and Society in England 1750-1950* (Bloomsbury Publishing 2019) 248-249.

[122] John Ramsay McCulloch, *Considerations on Partnerships with Limited Liability* (Longman, Brown, Green and Longmans 1856) 4-5, 11.

business community argued that the significant progress of the UK economy had been achieved under unlimited liability, and thus there was no economic incentive to alter the law.[123] This criticism is often attributed to the fact that contemporary businessmen were afraid of the competition that might have potentially come from joint-stock companies trading with limited liability.[124]

Another criticism of the introduction of general limited liability came from the judiciary. In particular, Lord Curriehill put forward an argument that limited responsibility of partners[125] was against natural justice and the obligation to pay debts.[126] Moreover, his Lordship argued that there was "no rule"[127] in UK law that would allow partners to be relieved from the liability for the debts of the partnership. At that time, however, various railway, canal, and mining companies operated with limited liability granted by Parliament under statutes because of the huge capital that was at stake in those companies.[128] Thus, the rule that shareholders' liability for the corporation's debts could be limited already existed in the law.[129] In addition, Lord Curriehill argued that limited liability would: (a) encourage fraud on creditors;[130] (b) increase the potential for excessive speculation[131] and; (c) create unfair competition for credit on the market, as some traders may, for various reasons, wish to opt out from limited liability, and thus end up in a worse business position.[132]

In sum, the argument against limited liability was based on the presumption that limited liability would be an unnecessary threat to public security due to its

[123] Sir Thomas Baring, HC Deb 27 June 1854, vol. 134 col. 685. See also The Royal Commission on Mercantile Law, First Report of the Commissioners Appointed to Inquire and Ascertain How Far the Mercantile Laws in the Different Parts of the United Kingdom of Great Britain and Ireland May Be Advantageously Assimilated and Also Whether Any and What Alterations and Amendments Should be Made in the Law of Partnership as regards the Question of the Limited or Unlimited Responsibility of Partners (1854) 7-8.

[124] McQueen (n 108) 81-86; James B Jefferys, 'Trends in Business Organization in Great Britain Since 1856, with Special Reference to the Financial Structure of Companies, the Mechanism of Investment and the Relations between the Shareholder and the Company' (PhD thesis, University of London, 1938) 41.

[125] As the major part of the nineteenth century debate concerned the introduction of partnerships *en commandite* into the law of the United Kingdom.

[126] The Royal Commission on Mercantile Law (n 123) 11.

[127] ibid.

[128] The privilege of limited liability was granted to these companies under the Chartered Companies Act 1837 and the Joint Stock Companies Registration and Regulation Act 1844. See Dabor (n 67) 110.

[129] ibid.

[130] The Royal Commission on Mercantile Law (n 123) 15-17.

[131] ibid 17-18.

[132] ibid 17.

potential to encourage speculation, harm creditors, and distort the market. On this account, it was postulated that its introduction would be detrimental to the entire nation.

Conversely, the argument of the advocates of limited liability was largely based on the economic rationale. In particular, it was argued that contemporary investors tended to invest their capital only in corporations, which operated with limited liability.[133] Likewise, those who would have to risk "losing their last shilling and their last acre"[134] in the course of business would simply not invest.[135] In this regard, The Times described the unlimited liability laws as having "a murderous character".[136] Similarly, The Economist referred to the contemporary state of the law as being, in principle and in practice, "bad".[137] In fact, there was evidence that the benefits conferred by limited liability upon shareholders encouraged numerous British companies to seek incorporation in the United States or in France.[138]

Moreover, it was argued that limited liability would protect innocent shareholders from the negative impact of fraudulent investors' activities, which was not the case under unlimited liability.[139] Furthermore, George Bramwell put forward an argument that the members of society should have a right to "the unrestrained and unfettered exercise of their own talents and industry", which, in Bramwell's view, was of greater importance for the community than the protection of contractual creditors.[140] This argument based on democratisation of capitalism was echoed by The Times, whose editors stated that limited liability would not only facilitate entrepreneurship, but would also encourage the middle class to engage in investing and wider commercial activities.[141]

Notably, involuntary creditors' protection was completely absent from the foregoing debate. Even though the law in the nineteenth century recognised that corporations could be held liable in tort, for example, a corporation could be sued in trover[142] or it could be held responsible for trespass,[143] the potential

133 Dabor (n 67) 99-100.
134 Robert Slaney, HC Deb 20 February 1851, vol. 114, col 846.
135 Report of the Select Committee on the Savings of the Middle and Working Classes HC 508 (6 June 1850) 109.
136 The Times, 28 July 1855.
137 The Economist Vol. XII, 1854, 698.
138 The Royal Commission on Mercantile Law (n 123) 239.
139 ibid 101.
140 ibid 23.
141 The Times, 27 July 1855, page 8. See also Kershaw (n 17) 24.
142 *Yarborough and Others v The Governor and Company of the Bank of England* (1812) 104 E.R. 991 at [991]-[993] per Lord Ellenborough C.J. A similar conclusion was reached by the court in *Duncan and Another v The Company of Proprietors of the Surrey Canal* (1821) 171 E.R. 763.
143 *Maund v The Monmouthshire Canal Company* (1842) 134 E.R. 186.

impact of the introduction of general limited liability on tort victims and other involuntary creditors was overlooked during the debate.[144] In turn, the debate centred around the economic aspects of the doctrine and its potential detrimental impact on contractual creditors.[145] For instance, in the Royal Commission's report, Lord Curriehill referred to "the temptation to embark on hazardous adventures"[146] encouraged by general limited liability. The risk of such activities, in his Lordship's opinion, would have to be borne by contractual creditors, as the potential for speculation would increase.[147] Unfortunately, Lord Curriehill did not consider the impact of 'hazardous adventures' on other types of creditors.

Arguably, in the nineteenth century, no one thought that the doctrine of limited liability could be extended in the future to shield the company's owners from the claims brought by third parties for the negligent acts of the company. This was because the concept of negligence, which is of crucial importance today, was not a fully developed legal principle at that time.[148] Furthermore, there was not a single case in the UK, where a company was rendered insolvent as a consequence of a tort claim.[149] As a result, involuntary creditors' protection was of considerably smaller importance at the time than contractual creditors' protection.[150]

Despite the lack of consensus on the question of whether general limited liability should be introduced into the law of the United Kingdom,[151] a sudden change took place and the law 'turned itself upside down' together with the introduction of general limited liability under the Limited Liability Act 1855 and the Joint Stock Companies Act 1856.[152] The reasons for this sudden reform and its unusually rapid pace are often attributed to various economic, social, and political

[144] Muscat (n 31) at 4.5.4; Pettet (n 98) 152.

[145] See in general, the comments of Lord Curriehill: The Royal Commission on Mercantile Law (n 123) 15-22. See also Robert A Bryer, 'The Mercantile Laws Commission of 1854 and the Political Economy of Limited Liability' (1997) 50(1) The Economic History Review 37, 46-48.

[146] The Royal Commission on Mercantile Law (n 123) 18.

[147] ibid 18.

[148] Muscat (n 31) at 4.5.4. In fact, the modern tort of negligence was not a fully developed legal doctrine until 1932, when this concept was expressly recognised by the judiciary in *Donoghue v Stevenson* [1932] AC 562. See Stefan HC Lo, *In Search of Corporate Accountability: Liabilities of Corporate Participants* (Cambridge Scholars Publishing 2016) 145.

[149] Muscat (n 31) at 4.5.4.

[150] ibid.

[151] The Royal Commissioners stated in their report that they were "embarrassed" by the contrariety of opinion on this matter: The Royal Commission on Mercantile Law (n 123) 5. In fact, three members of the Royal Commission ultimately voted in support of the doctrine of limited liability, whereas five members voted against its general introduction.

[152] Cottrell (n 98) 54.

factors, such as the wide support of the general public for limited liability,[153] the need of the government to find revenue to fund the Crimean War,[154] and the dominance of the liberal 'laissez-faire' approach in the contemporary political arena.[155] Accordingly, together with the enactment of these statutes, limited liability was no longer a privilege that was available only to a few members of society, but rather it became an easily available legal right.[156] In fact, during the debate on the Joint Stock Companies Bill, Robert Lowe argued that people should have liberty to deal with whom they wish without the unnecessary intervention of the state,[157] and that limited liability is "not a question of privilege; if anything, it is a right".[158]

What is of crucial importance, however, is that yet again neither the House of Commons nor the House of Lords considered the impact of the reform on involuntary creditors.[159] In turn, both Houses focused primarily on the risks posed by the introduction of general limited liability to contractual creditors.[160] The focus of Parliament on contractual creditors' protection is clearly evidenced by the words of Robert Lowe, who argued "in favour of human liberty" and in favour of the right of people to deal with whom they wish;[161] this emphasis on dealing indicates that the consequences that might arise from the application of the principle of limited liability to the contractual relationships between people were at the focal point of the Parliamentary discussion.[162] It can therefore be argued that the provisions of the new statutes were not intended to apply to involuntary creditors at all.[163]

It is worth noting that the new acts were not received well in all circles. For instance, for certain members of the House of Lords, it was "absurd"[164] that such an important change in the legislation was being passed so quickly. Indeed, the Bills were rushed through Parliament by the government at "almost indecent haste".[165] In addition, Edmund Phillips mocked the Limited Liability Bill, which, in his

[153] As Viscount Palmerston stated during the debate on the Limited Liability Bill in the House of Commons: "[Limited liability] is well understood by every man in the country (...)" HC Deb 26 July 1855, vol.139, col.1390.
[154] Mackie (n 28) 296-301.
[155] ibid 301-305; Lipton (n 65) 1292.
[156] Djelic (n 101) 599.
[157] HC Deb 1 February 1856, Vol.140, col.131.
[158] ibid col.129.
[159] Muscat (n 31) at 4.5.4; Mackie (n 28) 308.
[160] Mackie (n 28) 308.
[161] See Robert Lowe (n 157).
[162] Mackie (n 28) 308.
[163] ibid 295.
[164] Earl Grey, HL Deb 7 August 1855, Vol.139, col.1905.
[165] Laurence C B Gower, Ben Pettet and Daniel D Prentice, *Principles of Modern Company Law* (5th edn., Sweet and Maxwell 1992) 44-45; John Saville, 'Sleeping Partnership and Limited Liability' (1956) 8(3) Economic History Review 418, 430; Mackie (28) 299.

view, should have rather been called "An Act for the better enabling Adventurers to interfere with, and ruin, established traders, without risk to themselves".[166] Likewise, The Law Times described the Joint Stock Companies Act 1856 as being "monstrous in conception", undermining all safeguards against "fraud, folly and abuse".[167] Two years later, the same gazette referred to the 1856 Act as a "rogues' charter", which had proved to be a failure by allowing a man to avoid responsibility for his debts and liabilities.[168]

The discontent with the introduction of general limited liability could also be seen among some members of the public. For example, Gilbert and Sullivan wrote an operetta called *Utopia, Limited* where they satirised the idea of a limited liability company.[169] In the operetta, the king of a fictional state, Utopia, decrees that the entire kingdom, and all of its citizens, should become limited liability corporations in order to free themselves from any responsibilities. Notably, in Utopia, Gilbert and Sullivan question the appropriateness of the general application of limited liability by stating that its benefits are unjustified "unless accompanied by high standards of responsibility on the part of the corporate management".[170] Thus, the operetta argues for a wider protection of creditors, who, in the view of the authors, are prone to abuse under the doctrine.[171]

Nevertheless, the introduction of general limited liability in 1856 rendered the law of the United Kingdom as "the most permissive commercial law in Europe".[172] The main purpose of the reform was to stimulate economic growth by incentivising commercial activities.[173] Arguably, the new legal regime was

[166] Edmund Phillips, *Bank of England Charter, Currency, Limited Liability Companies, and Free Trade* (Richardson Brothers 1856) 36.

[167] 'The Law and the Lawyers' The Law Times, (London 26 July 1856) page 205. See also Djelic (n 101) 615.

[168] 'Anticipation and Experience' The Law Times (London 27 March 1854) 14. See Louis de Koker, 'Limited Liability Act of 1855' (2007) 26(5) The Company Lawyer 130, 130-131.

[169] William Schwenck Gilbert and Arthur Sullivan, Utopia Ltd (1893).

[170] Albert I Borowitz, 'Gilbert and Sullivan on Corporate Law' (1973) 59 American Bar Association Journal 1276, 1279.

[171] McQueen (n 108) 231.

[172] Cottrell (n 98) 41.

[173] Griffin (n 19) 99.

not intended to apply to involuntary creditors and their interests were not given adequate consideration during the debate.

C. *Salomon v Salomon* through the lens of involuntary creditors' protection

The judgment of the House of Lords in *Salomon v Salomon*[174] underpinned the development of the law on limited liability in the UK and the Commonwealth.[175] Because of its tremendous conceptual significance and vast practical implications, this decision has "stood the test of time"[176] and is considered today as "the key principle of company law".[177] Unfortunately, not much has been said about the wider impact of this decision on involuntary creditors.

In *Salomon*, the House of Lords took a literal approach towards statutory interpretation.[178] Because the language of the Companies Act 1862[179] only required seven persons associated for a lawful purpose to hold at least one share each in the company, the court ruled that the intention of the Parliament was to extend the benefit of incorporation (and thus of limited liability) to 'one-man' companies, as the statutory requirements were satisfied on the facts of the case.[180] The approach of the House of Lords was neatly summarised in the words of Lord Halsbury, who said that "the true intent and meaning of the Act" could only be derived from the Act itself.[181]

Notably, the House of Lords' judgment contrasts strikingly with the earlier decision of the Court of Appeal, where Lindley LJ held that "the legislature never contemplated an extension of limited liability to sole traders or to a fewer number

[174] *Salomon* (n 10).

[175] Susan Barber, *Company Law* (4th edn, Old Bailey Press 2003) 5; Hannigan (n 10) 42; Ross Grantham and Charles Rickett, *'The Bootmaker's Legacy to Company Law Doctrine' in Ross Grantham and Charles Rickett (eds), Corporate Personality in the 20th Century* (Hart Publishing 1998) 1; Lipton (n 33) 453-454; Eneless Nyoni and Tina Hart, 'The Concept of Limited Liability and the Plight of Creditors within Corporate Governance and Company Law: A UK Perspective' (2018) 5(2) INTERULAWEAST - Journal for International and European Law, Economics and Market Integrations 309, 312.

[176] Davies and Worthington (n 17) 200.

[177] Lord Cooke of Thorndon, *The Hamlyn Lectures: Turning Points of the Common Law* (Sweet & Maxwell 1997) 17.

[178] Roman Tomasic, Stephen Bottomley and Rob McQueen, *Corporations Law in Australia* (Federation Press 2002) 33-34. Ernest Lim, 'Of "Landmark" or "Leading" Cases: Salomon's Challenge' (2014) 41(4) Journal of Law and Society 523, 534: Lipton (n 33) 469.

[179] Ss. 6 and 8 in particular.

[180] *Salomon* (n 10) at [29]-[33] per Lord Halsbury, at [37]-[40] per Lord Watson, at [45]-[47] per Lord Herschell, at [51] per Lord Macnaghten, at [54] per Lord Davey. See Lim (n 178) 534.

[181] *Salomon* (n 10) at [31] per Lord Halsbury.

than seven",[182] and although in the *Salomon* case there were seven members, "six of them were members simply in order to enable the seventh himself to carry on business with limited liability".[183] Ultimately, his Lordship held that the sole intention of Mr Salomon was to obtain the benefit of limited liability to "defraud creditors".[184] Lindley LJ's approach was, however, firmly rejected by the House of Lords and the Court of Appeal's decision was unanimously reversed.[185]

It is argued that in *Salomon*, the application of the literal approach led to absurdity.[186] Namely, it is highly doubtful that the true intention of the Parliament was to extend the benefit of limited liability to one-man companies.[187] In fact, it is evident that Parliament's view was that the benefit of limited liability was available only to large businesses, and not to small private companies or sole traders.[188] Therefore, in the opinion of one commentator, in rejecting the clear meaning of the statute, the House of Lords showed severe "jurisprudential ineptitude".[189]

Moreover, Kahn-Freund described the *Salomon* judgment as "calamitous"[190] on the ground that the House of Lords allowed a sole trader, or groups of traders, to carry out business through a limited liability company in cases where there is no need for outside capital and where no specific business risk is involved. Therefore, in such cases, the economic benefits of limited liability are lost and the application of the doctrine seems unjustified. In addition, according to Kahn-Freund, the law has failed to give adequate protection to creditors, which should be the corollary of

[182] *Broderip v Salomon* [1895] 2 Ch. 323 at [337].

[183] ibid.

[184] ibid at [339]. Lopes LJ said in this regard that "it would be lamentable if a scheme like this could not be defeated": ibid at [340]-[341].

[185] See, for example, *Salomon* (n 10) at [31] per Lord Halsbury, at [38] per Lord Watson, at [45]-[46] per Lord Herschell, at [51]-[52] per Lord Macnaghten. See also Lim (n 178) 532.

[186] Ross Charnock, 'Lexical Indeterminacy: Contextualism and Rule-Following' in Anne Wagner, Wouter Werner and Deborah Cao (eds), *Interpretation, Law and the Construction of Meaning. Collected Papers on Legal Interpretation in Theory, Adjudication and Political Practice* (Springer 2006) 23-24.

[187] ibid.

[188] Paddy Ireland, 'The Rise of the Limited Liability Company' (1984) 12 International Journal of the Sociology of Law 239, 241-244; Lim (n 178) 534.

[189] Patrick F Higgins, *The Law of Partnership in Australia and New Zealand* (Sydney, Law Book Company of Australasia 1963) 16. See also Paul Halpern, Michael Trebilcock, and Stuart Turnbull, 'An Economic Analysis of Limited Liability in Corporation Law' (1980) 30(2) The University of Toronto Law Journal 117, 119.

[190] Otto Kahn-Freund, 'Some Reflections on Company Law Reform' (1944) 7.1/2 The Modern Law Review 54, 54-55. See also Simon Goulding, *Principles of Company Law* (Cavendish Publishing Limited 1996) 49; Gonzalo Villalta Puig, 'A Two-Edged Sword: Salomon and the Separate Legal Entity Doctrine' (2000) 7.3 Murdoch University Electronic Journal of Law 1, 17-18. <www5.austlii.edu.au/au/journals/MurdochUeJlLaw/2000/32.html> accessed 1 August 2020.

limited liability.[191] Indeed, creditors have to bear the risks associated with dealing with limited liability corporations.[192] While large contractual creditors, such as banks, are able to protect themselves against said risks, tort creditors and other types of involuntary creditors cannot.[193]

In justifying the impact of the doctrine of limited liability on unsecured creditors, Lord Macnaghten said that Salomon Ltd.'s creditors "may be entitled to sympathy, but they have only themselves to blame for their misfortunes".[194] This is because they had a long relationship with Mr Salomon and they "had full notice"[195] that they were dealing with a company rather than with an individual. In the modern market reality, however, it is questionable whether Lord Macnaghten's statement is still relevant. In particular, it is doubtful whether involuntary creditors, especially tort victims, ever had any notice that they were dealing with a company before they were injured. It is therefore highly questionable that tort creditors have "themselves to blame".[196] Hence, it can be argued that the principle of limited liability should have never been extended to involuntary creditors, as neither the judiciary, nor the Parliament had such an extension of the application of the doctrine in mind.[197] Instead, in creating the limited liability regime, the intention of the lawmakers was to encourage people to freely engage in commercial activities.[198] Indeed, Lord Macnaghten mentioned Salomon Ltd.'s unsecured creditors, who had notice that they were dealing with a limited liability company;[199] they had

[191] Kahn-Freund (n 190) 55. This argument was echoed by Scanlan, who stated that "the law needs to go further in providing a means of protection for the unsecured creditor of private limited companies" Gary Scanlan, 'The Salomon Principle' (2004) 25.7 The Company Lawyer 196, 198.

[192] Puig (n 190) 19.

[193] ibid; Lo (n 18) 121; Lipton (n 33) 482.

[194] *Salomon* (n 10) at [53].

[195] ibid.

[196] Ewan McGaughey, 'Donoghue v Salomon in the High Court' (2011) 4 Journal of Personal Injury Law 249, 253.

[197] ibid; Lo (n 148) 146. The problems associated with the application of the limited liability principle to torts were already noticed by Edward Cox, who wrote that, in practice, "no person will bring an action (in tort) against a company from which he can recover nothing". Edward Cox, *The Law and Practice of Joint Stock Companies* (London, Law Times Office 1856) 5.

[198] Griffin (n 19) 99.

[199] *Salomon* (n 10) at [53].

therefore chosen to enter into a commercial relationship with it.[200] Likewise, in the Parliament, Robert Lowe argued for the "right of people"[201] to choose with whom they wish to deal. Again, therefore, the emphasis on the ability to choose with whom one wanted to enter into commercial relationships was of paramount importance. Involuntary creditors, however, cannot choose with whom they want to deal;[202] they cannot bargain with the corporation,[203] and thus they cannot protect themselves from the risks associated with the application of the doctrine of limited liability.[204]

D. Interim conclusion

In sum, today the doctrine of limited liability transfers business risks from investors to involuntary creditors without compensation.[205] This, in turn, incentivises corporate recklessness, as the corporate veil protects investors from any liabilities arising from the company's activity.[206] This problem is exacerbated by multinational corporate groups, as parent companies often use the benefit of limited liability for the sole purpose of avoiding responsibility for the wrongs done

[200] It is argued that unsecured creditors "bargain with the corporation", see Thomas H Jackson, 'Bankrupcy, Non-Bankrupcy Entitlements, and the Creditors' Bargain' in Richard A Posner (ed), *Corporate Bankruptcy: Economic and Legal Perspectives* (Cambridge University Press 1996). For a well-thought critique of unsecured creditors' bargain, see Lynn M LoPucki, 'The Unsecured Creditor's Bargain' (1994) 80(8) Virginia Law Review 1887. For a response to LoPucki's arguments, see Susan Block-Lieb, 'The Unsecured Creditor's Bargain: A Reply' (1994) 80(8) Virginia Law Review 1989.

[201] See Robert Lowe (n 157).

[202] McGaughey (n 196) 253-254.

[203] Leebron (n 46) 1639-40; LoPucki (n 200) 1897-1898.

[204] Lo (n 148) 145-146.

[205] Janet Cooper Alexander, 'Unlimited Shareholder Liability Through A Procedural Lens' (1992) 106 Harvard Law Review 387, 390; Lo (n 18) 121; Puig (n 190) 19.

[206] Hansmann and Kraakman (n 32) 1920; Ribstein (n 28) 81; Leebron (n 46) 1565; Price (n 46) 441-442; Ireland (n 49) 838; Muchlinski (n 39) 915-916; Lipton (n 33) 480-481.

by their subsidiaries.[207] These are the consequences of limited liability, which, as the foregoing discussion has shown, were not foreseen by its inventors.[208]

III. Involuntary Creditors' Protection Under the Modern Doctrine of Limited Liability

A. The economic rationale behind limited liability

The foregoing analysis evidenced that the doctrine of limited liability was introduced in the UK with the purpose of stimulating economic growth. This analysis also showed that the interests of involuntary creditors were not given adequate consideration when the doctrine was introduced. Subsequently, the principle has been extended to shield a company's owners from liabilities towards this particular group of creditors, which was not the intention of the lawmakers.

That being said, limited liability is considered today as a "birth right"[209] of a corporation as well as its most attractive feature.[210] In fact, in modern times, most companies decide to conduct their businesses under limited liability.[211] For instance, from 2004, limited companies have consistently accounted for over 96% of all corporate bodies in the UK.[212] In contrast, between 2017 and 2018, only four thousand three hundred seventy-four companies traded in Britain with unlimited

[207] See the comments of Templeman LJ in *Re Southard & Co. Ltd* [1979] 1 W.L.R. 1198, who stated at [1208] that "A parent company may spawn a number of subsidiary companies, all controlled directly or indirectly by the shareholders of the parent company. If one of the subsidiary companies, to change the metaphor, turns out to be the runt of the litter and declines into insolvency to the dismay of its creditors, the parent company and the other subsidiary companies may prosper to the joy of the shareholders without any liability for the debts of the insolvent subsidiary". See also Rühmkorf (n 40) 183-184; Peter Muchlinski, 'Holding Multinationals to Account: Recent Developments in English Litigation and the Company Law Review' (2002) 23 The Company Lawyer 168, 168-169.

[208] It is argued that, in the modern times, "limited liability has been carried unthinkingly beyond the original objective of insulating the ultimate investor from the debts of the enterprise": Blumberg (n 12) 575.

[209] Leebron (n 46) 1569.

[210] Harry G Henn, *Handbook of the Law of Corporations and Other Business Enterprises* (West Publishing Company 1970) 96; David H Fater, *Essentials of Corporate and Capital Formation* (John Wiley & Sons 2009) 16.

[211] Ali Imanalin, 'Rethinking Limited Liability' (2011) 7 Cambridge Law Review 89, 89-90.

[212] Companies House, 'Official Statistics: Companies register activities: 2019 to 2020' (25 June 2020) <www.gov.uk/government/publications/companies-register-activities-statistical-release-2019-to-2020/companies-register-activities-2019-to-2020> accessed 13 July 2020.

liability.[213] Thus, it is clear that a limited liability company is the most popular business vehicle in the UK.

One of the main reasons for the prominence of the doctrine of limited liability and its widespread acclamation[214] is the fact that it stimulates economic growth by incentivising business endeavours.[215] Namely, because there is a linear relationship between returns on a particular asset and its systematic risks, high returns from an investment cannot be achieved without accepting the substantial risk of potential financial losses.[216] Thus, because investors tend to be risk averse,[217] they may often feel discouraged from putting their personal assets at stake in a risky business venture, as these assets are not protected from the claims raised by the company's creditors. In this regard, the doctrine of limited liability provides that the liability of the company's members is restricted to the amount invested in the company.[218] Consequently, the shareholders will not be held personally liable for a sum greater than what they have already invested; their personal assets will remain shielded from the claims of the company's creditors.[219] Effectively, the doctrine of limited liability shifts the risk of potential financial losses away from shareholders and places it upon creditors.[220] As a result, investors can 'sleep more easily at night' knowing that limited liability protects them from the risk of bankruptcy.[221] Also, the principle of limited liability allows shareholders to diversify their portfolios

213 Companies House, 'Official Statistics: Companies register activities: 2017 to 2018' (28 June 2018) <www.gov.uk/government/publications/companies-register-activities-statistical-release-2017-to-2018/companies-register-activities-2017-to-2018> accessed 13 July 2020.

214 See Bernard F Cataldo, 'Limited Liability with One Man Companies and Subsidiary Corporations' (1953) 18(4) Law and Contemporary Problems 473, 473-474.

215 Davies and Worthington (n 17) 191-192; Griffin (n 19) 99.

216 Ronald J Gilson, 'Value Creation by Business Lawyers: Legal Skills and Asset Pricing' (1984) 94 Yale Law Journal 239, 313.

217 Bainbridge and Henderson (n 21) 47. For a thorough analysis of risk aversion of human beings, see Elke U Weber, Ann-Renee Blais and Nancy E Betz, 'A Domain-Specific Risk - Attitude Scale: Measuring Risk Perceptions and Risk Behaviors' (2002) 15(4) Journal of Behavioral Decision Making 263, 263-290. See also Laura Concina, Risk Attitude & Economics (FonCSI 2014) 12-16.

218 Davies and Worthington (n 17) 191.

219 Davies (n 13) 60; Bae Kim (n 16) 73; Peterson (n 22) 63.

220 Lewis D Solomon et al., *Corporation Law and Policy* (2nd edn., West Pub. Co 1988) 242; Bainbridge and Henderson (n 21) 47; Millon (n 30) 1355.

221 William Reader, 'Versatility Unlimited: Reflections on the History and Nature of the Limited Liability Company' in Tony Orhnial (ed), *Limited Liability and the Corporation* (London: Croom Helm 1982) 191.

more easily, which reduces the shareholders' company-specific risks, and thus their potential financial losses.[222]

Moreover, the doctrine diminishes monitoring costs that investors would otherwise have to incur to monitor managerial behaviour.[223] Namely, directors may sometimes act in a manner detrimental to the interests of shareholders.[224] When shareholders' personal liability is not limited, there is a risk that they would have to bear the costs of the directors' course of conduct.[225] For that reason, shareholders would have to actively monitor directors' decision-making, which is a costly, complicated, and time-consuming process.[226] Accordingly, the doctrine of limited liability eliminates shareholders' personal liability for the company's debts, and thus active monitoring is no longer necessary.[227] Likewise, limited liability diminishes the costs of monitoring other shareholders.[228] Namely, when the liability of the company's members is not limited, the obligation to pay off the company's debts could be disproportionally imposed upon the wealthier shareholders.[229] As a result, the company's members would have to monitor each other to anticipate potential liabilities.[230] Under the doctrine of limited liability this problem ceases to exist, as the identity of other shareholders becomes irrelevant.[231]

In addition to the benefits mentioned above, limited liability incentivises managerial efficiency,[232] facilitates optimal investment decision-making,[233] encourages public investment,[234] and enables smooth transferability of shares on the market.[235] To put it briefly, the doctrine of limited liability incentivises investments that would not otherwise take place.[236] Given its salient economic role, limited

[222] Davies and Worthington (n 17) 192; Easterbrook and Fischel (n 15) 96; Price (n 46) 448-449.
[223] Easterbrook and Fischel, 'Limited Liability and the Corporation' (n 15) 94-95.
[224] Kershaw (n 17) 25.
[225] ibid.
[226] Millon (n 30) 1312-1313.
[227] ibid. It is worth noting here that oftentimes diminished monitoring costs open up a path for directorial fraud, see Burcu S Avci, Cindy A Schipani, and H Nejat Seyhun. 'Do Independent Directors Curb Financial Fraud: The Evidence and Proposals for Further Reform' (2018) 93 Indiana Law Journal 757.
[228] Easterbrook and Fischel (n 15) 95. See also Halpern, Trebilcock, and Turnbull (n 189).
[229] Frank H Easterbrook and Daniel R Fischel, *The Economic Structure of Corporate Law* (Harvard University Press 1996) 42.
[230] ibid.
[231] ibid.
[232] ibid.
[233] ibid 43-44.
[234] Davies (n 13) 63-65.
[235] Price (n 46) 444-445; Millon (n 30) 1313.
[236] Millon (n 30) 1312.

liability is rightly regarded as the "distinguishing feature" of corporate law,[237] which "has made much of our economic progress possible".[238] Consequently, limited liability forms an inextricable part of modern capitalism and modern society.

Despite the plethora of economic benefits induced by the doctrine of limited liability, it is evident that the principle also has its costs.[239] As this paper has already discussed, the doctrine allows companies to easily externalise their business risks and incentivises corporate recklessness.[240] Consequently, the doctrine puts involuntary creditors in danger, as it shifts risks away from shareholders onto involuntary creditors without compensation.[241] It is therefore worth considering how the modern law protects involuntary creditors from the negative impact of the application of the doctrine of limited liability, and whether these protection mechanisms are effective.

B. Modern involuntary creditors' protection mechanisms

(i) Piercing the Corporate Veil

In certain instances, the court may 'pierce the corporate veil', and may thus hold the company's shareholders liable for the debts of the corporation by denying that a company is a separate person in the eyes of the law.[242] Traditionally, UK courts were willing to 'pierce the veil' in "appropriate circumstances".[243] For instance, the veil will be pierced when a statute so permits, e.g., directors can be found personally liable for wrongful trading.[244] Likewise, anyone who knowingly

[237] Easterbrook and Fischel, *The Economic Structure of Corporate Law* (n 229) 40.

[238] John Hicks, 'Limited Liability: the Pros and Cons' in Tony Orhnial (ed), *Limited Liability and the Corporation* (London: Croom Helm 1982) 12.

[239] Peterson (n 22) 64.

[240] Hansmann and Kraakman (n 32) 1920; Muchlinski (n 39) 915-916.

[241] Lo (n 18) 121.

[242] *Prest v Petrodel Resources* [2013] 2 A.C. 415 at [16] per Lord Sumption. Davies and Worthington (n 17) 197-198; Dmitry Vlasov, 'Liability of a Puppeteer for a Puppet: a Recent Development in Law on Piercing the Corporate Veil' (2012) 33 The Company Lawyer 356, 356; Andrew Bowen, 'Concealment, Evasion and Piercing the Corporate Veil: Prest v Petrodel Resources Ltd' Business Law Bulletin (April 2014) 1.

[243] *Customs and Excise Commissioners v Hare* [1996] 2 B.C.L.C. 500 at [511] per Rose LJ.

[244] Insolvency Act 1986, s. 214 as amended; *Re Produce Marketing Ltd* [1989] BCLC 520; *Re Continental Assurance Company of London Plc (Signer v. Beckett)* [2007] B.C.L.C. 287.

contributed to the fraudulent business conduct of the company can be held personally liable for fraudulent trading.[245]

The corporate veil can also be pierced by the courts under common law principles. Namely, under the law of agency, owners of a company can be found liable for the company's actions, provided that the actions were within the scope of actual[246] or ostensible[247] authority.[248] For instance, in *Smith Stone & Knight Ltd v Birmingham Corporation,* Atkinson J. held that the subsidiary company was so closely connected with the parent company that it was in fact operating as the parent's agent.[249] Normally, however the presumption of an agency relationship is difficult to establish without an express agreement between the interested parties.[250] Indeed, in *Adams v Cape Industries*, the argument based on agency failed completely, as no contractual relationship that could amount to a parent-agent relationship was found between the parties even though the subsidiaries were wholly owned by the parent.[251] Such cases are therefore very fact-sensitive.[252]

In the context of corporate groups, in *DHN Food Distributors Ltd v Tower Hamlets London Borough Council*,[253] Lord Denning advanced the "single economic unit" argument, which initially was seen as undermining the rigidity of the *Salomon* judgment.[254] In this case, Lord Denning, with whom Lords Goff[255] and

[245] Insolvency Act 1986, s. 213 as amended. For example, in *Re Gerald Cooper Chemicals Limited (in Liquidation)* [1978] 1Ch 262 it was the creditor, who was found liable for breaching s. 213.

[246] *Hely-Hutchinson v Brayhead Ltd* [1968] 1 Q.B. 549 at [583] per Lord Denning. See, in general, Peter Watts and Francis M B Reynolds, *Bowstead and Reynolds on Agency* (21st edn., Sweet & Maxwell 2018) at 3.001-3.006.

[247] *Freeman & Lockyer (A Firm) v Buckhurst Park Properties (Mangal) Ltd. and Another* [1964] 2 Q.B. 480 at [503]-[506] per Lord Diplock; *Hely-Hutchinson* (n 246) at [583] per Lord Denning. See also Kershaw (n 17) 112-120.

[248] On agency law in general, see Davies and Worthington (n 17) 149-191.

[249] *Smith Stone & Knight Ltd v Birmingham Corporation* [1939] 4 All E.R. 116 at [121]-[122].

[250] *Southern v Watson* [1940] 3 All E.R. 439; Davies and Worthington (n 17) 203.

[251] *Adams* (n 43) at [545]-[550] per Slade L.J; Kershaw (n 17) 58; Davies and Worthington (n 17) 203.

[252] Atkinson J himself admitted that the question in *Smith Stone & Knight Ltd* was one of fact: *Smith Stone & Knight Ltd* (n 249) at [121]. In this regard, Kerr LJ stated in *JH Rayner (Mincing Lane) Ltd v Department of Trade & Industry* [1989] Ch 72 (CA) at [190] that the facts of *Smith Stone & Knight* were so unique that "no conclusion of principle could be derived from that case".

[253] *DHN Food Distributors Ltd v Tower Hamlets London Borough Council* [1976] 1 WLR 852 (CA).

[254] David Sugarman and Frank Webb, 'Three-in-One: Trusts, Licences and Veils' (1977) 93 Law Quarterly Review 170, 175.

[255] Lord Goff's reasoning was based on the facts that the subsidiaries were wholly owned by the parent company and that that there was no business activity in them: *DHN* (n 253) at [860]-[866]; Kershaw (n 17) 63.

Shaw[256] agreed, held that because the parent company wholly owned shares in the subsidiaries and had complete control over them, justice required that the three companies should be treated as one; they constituted one economic entity.[257] However, this principle has subsequently been considerably qualified. In *Woolfson v Strathclyde Regional Council*, Lord Keith distinguished *DHN* on the facts and held that the court in *DHN* did not apply the law correctly, as the corporate veil could only be pierced in special circumstances indicating the existence of "a mere façade concealing the true facts".[258] Moreover, in *Adams*, Slade L.J. held, citing Roskill L.J. in *Albacruz (Cargo Owners) v Albazero*,[259] that it is a fundamental principle of the law that each company in the group of companies is a separate legal person.[260] Accordingly, the court is "not free" to disregard the *Salomon* principle merely because justice so requires.[261] For those reasons, even though the 'single economic unit' principle has never been completely rejected,[262] the judgment in *DHN* is often seen as an "aberration",[263] which is strictly confined to its facts.[264] Therefore, the 'single economic unit argument' is of little practical relevance today.[265] From the perspective of involuntary creditors this is unfortunate, because, had Lord

[256] Shaw LJ focused on the fact that the subsidiary did not engage in any kind of trading and had no real business: *DHN* (n 253) at [866]-[868].

[257] ibid at [860]. Thomas K Cheng, 'The Corporate Veil Doctrine Revisited: A Comparative Study of the English and the US Corporate Veil Doctrines' (2011) 34 Boston College International and Comparative Law Review 329, 331-332.

[258] *Woolfson v Strathclyde Regional Council* [1978] S.C. (H.L.) 90 at [95]-[96].

[259] *Albacruz (Cargo Owners) v Albazero* [1977] AC 774 at [807]: "Each company in a group of companies is a separate legal entity possessed of separate legal rights and liabilities so that the rights of one company in a group cannot be exercised by another company in that group even though the ultimate benefit of the exercise of those rights would ensure beneficially to the same person or corporate body irrespective of the person or body in whom those rights were vested in law".

[260] *Adams* (n 43) at [532].

[261] ibid [536]-[537]. See also the comments of Munby J in *Ben Hashem v Ali Shayif* [2009] EWHC 864 (Fam) at [160].

[262] Kershaw (n 17) 69-70.

[263] F G Rixon, 'Lifting the Veil Between Holding and Subsidiary Companies' (1986) 102 Law Quarterly Review 415, 422.

[264] Ernest Lim, *Sustainability and Corporate Mechanisms in Asia* (Cambridge University Press 2020) 247.

[265] See, for instance, the case of *Gripple Ltd v Revenue and Customers Commissioners* [2010] EWHC 1609 (Ch) at [23]-[24], where Henderson J firmly rejected the "single economic unit" argument put forward by the counsel for the appellant and described such an argument as "an adventurous submission". See also the comments of Flaux J in *Linsen International Ltd and others v Humpuss Sea Transport Pte Ltd and others* [2012] Bus. L.R. 1649 at [19], [39], [58] and [126].

Denning's argument been accepted by the courts, it would arguably have given involuntary creditors a considerable degree of protection.

Furthermore, separate legal personality of a company can be disregarded in cases where the company was a "sham",[266] a "cloak",[267] a "mask",[268] a "puppet",[269] or a "mere façade",[270] which was used solely as a stratagem to conceal the true facts, such as the existence of fraud, or to evade existing liabilities.[271] The above epithets are considered as synonymous.[272] Nevertheless, what they really mean is incredibly unclear.[273] In this regard, in *Adams*, the court stated that the authorities give "rather sparse guidance" as to what kind of arrangement amounts to a "façade".[274] It did not, however, provide any clarification on this issue.[275]

In *Prest v Petrodel Resources*, which is the leading case on piercing the veil in the UK, Lord Sumption stated, having reviewed the authorities on this issue, that piercing the veil is an applicable remedy only in cases where a person tries to deliberately evade existing legal obligations, or liabilities, through relying on the corporate form for an improper purpose.[276] His Lordship also held that, in numerous instances, an evident legal relationship exists between the company and its owners, which makes it unnecessary to pierce the corporate veil at all; the same outcome can be reached through "more conventional"[277] methods, such as the law of agency. Therefore, the veil will be pierced only when doing so is absolutely necessary.[278] Consequently, following *Prest*, the scope of the doctrine has become

[266] *Gilford Motor Company, Limited v Horne* [1933] Ch. 935 at [961] per Lord Hanworth MR, at [965] per Lawrence LJ, at [969] per Romer LJ.

[267] ibid at [961] per Lord Hanworth MR, at [965] per Lawrence LJ, at [969] per Romer LJ.

[268] *Jones v Lipman* [1962] 1 WLR 832 at [836] per Russell J.

[269] *Wallersteiner v Moir* [1974] 1 W.L.R. 991 at [1013] per Denning MR.

[270] *Woolfson v Strathclyde Regional Council* [1978] S.C. 90 (HL) at [96] per Lord Keith with whom Lord Wilberforce, Lord Fraser of Tullybelton and Lord Russell of Killowen agreed.

[271] ibid; *Trustor AB v Smallbone and others (No 2)* [2001] 1 W.L.R. 1177 at [20] per Sir Andrew Morritt V-C. See also Paweł Słup, 'Piercing the Corporate Veil – A Common Pattern?' (2018) 24.1 Comparative Law Review 287, 298.

[272] *Ben Hashem* (n 261) at [150] Mr Justice Munby.

[273] Davies and Worthington (n 17) 202.

[274] *Adams* (n 43) at [543] per Slade LJ; Davies and Worthington (n 17) 202.

[275] *Adams* (n 43) at [543] per Slade LJ.

[276] *Prest* (n 242) at [35].

[277] ibid. Lord Neuberger agreed with Lord Sumption on this issue, see para [62].

[278] ibid at [35] per Lord Sumption; Pey Woan Lee, 'The Enigma of Veil Piercing' (2015) 26(1) International Company and Commercial Law Review 28, 33.

extremely narrow, and piercing the veil can be regarded as a "remedy of last resort".[279]

The foregoing suggests that piercing the corporate veil might not be an appropriate remedy for involuntary creditors. Namely, the above discussion evidenced that courts are very reluctant to disregard the corporate form, and will only do so in "exceptional" circumstances.[280] In fact, following *Prest*, the veil will only be pierced when the corporate form is used to evade existing legal obligations;[281] the scope of the doctrine has thus become very narrow.[282] Therefore, because future torts cannot be regarded as existing obligations, involuntary creditors will not be able to request the court to disregard the corporate form and hold the company's owners liable for the wrong done by the company or its subsidiary.[283] Moreover, piercing the veil involves a considerable degree of judicial discretion and it is hard to anticipate *ex ante* whether the veil will be pierced, which leads to uncertainties and increases the costs of fact-specific litigation.[284] For the above reasons, the veil piercing doctrine is a "remedy of last resort",[285] which is rarely successful. Even though the single economic unit argument could initially be seen as a light at the end of the tunnel for involuntary creditors, the importance of this principle has been considerably limited. On this account, the veil piercing doctrine does not allow even the most deserving victims, such as the employees who contracted asbestosis in the case of *Adams*, to be duly compensated for their losses.[286]

(ii) Bypassing Limited Liability under Tort Law - Chandler v Cape Plc

Holding a parent company directly liable for the harm caused by its subsidiary is an alternative approach to the issue of involuntary creditors' protection available under tort law. Namely, in *Chandler v Cape Plc*, Wyn Williams J

279 Lee (n 278) 33; Alexander Schall, 'The New Law of Piercing the Corporate Veil in the UK' (2016) 13(4) European Company and Financial Law Review 549, 555-556; Edwin Mujih, 'Piercing the Corporate Veil as a Remedy of Last Resort after Prest v Petrodel Resources Ltd: Inching towards Abolition?' (2016) 37(2) The Company Lawyer 39, 49.

280 Muchlinski (n 39) 919; Ernest Lim, 'Salomon Reigns' (2013) 129 Law Quarterly Review 480, 483; Petrin and Choudhury (n 39) 775.

281 *Prest* (n 242) at [34]-[35] per Lord Sumption, at [81] per Lord Neuberger.

282 Lee (n 278) 30; Schall (n 279) 552-553; Stefan H C Lo, 'Piercing of the Corporate Veil for Evasion of Tort Obligations' (2017) 46.1 Common Law World Review 42, 45.

283 Lipton (n 33) 480; Phillip Morgan, 'Vicarious Liability for Group Companies' (2015) 31 Journal of Professional Negligence 276, 283; Lo (n 282) 44; Christopher Arvidsson, 'The Piercing Doctrine: Re-examining Evasion' (2019) 40(10) The Company Lawyer 320, 322.

284 Muchlinski (n 39) 923.

285 Schall (n 279) 555-556.

286 For a similar argument, see Imanalin (n 211) 90.

held, relying on the tripartite test established in *Caparo v Dickman*,[287] that the parent company owed a duty of care to the employee, who contracted asbestosis in the course of his employment with the subsidiary.[288] Following *Caparo*, a duty of care arises provided that (a) the harm suffered is reasonably foreseeable; (b) there is a high degree of "proximity" between the parties and; (c) the imposition of the duty is "fair, just and reasonable".[289] Accordingly, in *Chandler*, the court found that the harm was reasonably foreseeable, as the parent company had knowledge of the working conditions of the subsidiary's employees.[290] Moreover, the court found that, because the parent company dictated policy in regard to health and safety measures of the subsidiary, the parent retained responsibility for the safety of the subsidiary's employees.[291] Therefore, a degree of proximity existed between the parties.[292] Furthermore, the court firmly admitted that imposing a duty of care in relation to exposure to asbestos was fair, just and reasonable.[293] Ultimately, it was held that the parent company owed a duty of care to the subsidiary's employees. This judgment was subsequently affirmed on appeal.[294]

In its decision, the Court of Appeal elaborated on Mr Justice Wyn Williams's approach and put forward a four-part test, under which the law can impose responsibility for the safety of the subsidiary's employees upon the parent company.[295] Accordingly, once it is established that the parent is in "relevant control of the subsidiary's business",[296] the parent's responsibility will be assumed when: (a) the business of the parent and the subsidiary are in a relevant respect the same; (b) the parent has "superior knowledge" of the health and safety standards in the relevant industry; (c) the parent has superior knowledge of the subsidiary's working conditions and; (d) "the parent knows or ought to have foreseen that the subsidiary

[287] *Caparo v Dickman* [1990] 2 A.C. 605 at [617]-[618] per Lord Bridge of Harwich.

[288] [2011] EWHC 951 (QB) at [77] per Mr Justice Wyn Williams. For a detailed analysis of this judgment, see Martin Petrin, 'Assumption of Responsibility in Corporate Groups: Chandler v Cape plc.' (2013) 76(3) The Modern Law Review 603, 607-609.

[289] *Caparo* (n 287) at [617]-[618] per Lord Bridge of Harwich.

[290] *Chandler* (n 288) at [73]-[74] per Mr Justice Wyn Williams.

[291] ibid at [71] and [75] per Mr Justice Wyn Williams.

[292] ibid at [75] per Mr Justice Wyn Williams.

[293] ibid at [76] per Mr Justice Wyn Williams.

[294] *Chandler v Cape Plc* [2012] 1 W.L.R. 3111 (CA).

[295] ibid at [80] per Arden LJ with whom Moses LJ and McFarlane LJ agreed. See Rühmkorf (n 40) 175-176.

[296] *Chandler* (n 294) at [46] per Arden LJ. See also Petrin (n 288) 610-611.

or its employees would rely on its using that superior knowledge for the employees' protection".[297]

Because of its wide implications, the judgment in *Chandler* marked an important step in holding the parent company liable for the injury sustained by its subsidiary's employees in the course of their employment.[298] Accordingly, it gives involuntary creditors a possibility to sue the parent company under tort law, which effectively bypasses the separate legal personality of the company and circumvents the doctrine of limited liability. Therefore, following *Chandler*, things might have changed, and the tide of the conflict between tort law and company law, wherein company law for long took the upper hand, might have turned as well.[299]

However, *Chandler* does not in fact establish any clear principle according to which future cases could be decided.[300] Namely, it is unclear what the Court of Appeal meant by 'relevant control'; it is unclear how much 'control' is sufficient for responsibility to be assumed.[301] Likewise, under this approach, it is arguably not necessary to show that the parent company exercised control over the subsidiary's health and safety standards.[302] In turn, showing that the parent company had 'a general practice' of intervening in the management of the subsidiary will be enough to satisfy the test.[303] In addition, as Petrin and Choudhury argue, this approach is both underinclusive and overinclusive.[304] It is underinclusive because establishing that the parent company failed to exercise "relevant control" should not automatically exempt the parent from incurring liability.[305] It is overinclusive because virtually all corporate groups have certain uniform group policies, and thus practically every parent company will satisfy the 'control' requirement.[306] Moreover, following *Chandler*, the existence of the duty of care will be highly dependent on the nature of the relationship between the parent company and the subsidiary; such a relationship must be sufficiently close.[307] For example, in *Thompson v The Renwick Group Plc*, it was held that the fact that the parent company merely held shares in the subsidiary was not enough to impose liability under tort

[297] ibid at [80] per Arden LJ.

[298] Rühmkorf (n 40) 176-177; Julian Fulbrook, 'Chandler v Cape Plc: Personal Injury: Liability: Negligence' (2012) 3 Journal of Personal Injury Law C135-C139.

[299] McGaughey (n 196) 249.

[300] Petrin and Choudhury (n 39) 778.

[301] ibid; Petrin (n 288) 612.

[302] Petrin (n 288) 613.

[303] *Chandler* (n 294) at [80] per Arden L.J; Petrin (n 288) 613; Petrin and Choudhury (n 39) 778.

[304] Petrin and Choudhury (n 39) 778.

[305] ibid.

[306] ibid.

[307] Peter Rott, 'Directors' Duties and Corporate Social Responsibility under German Law - Is Tort Law Litigation Changing the Picture' (2017) 1 Nordic Journal of Commercial Law 11, 22.

law upon the parent.[308] Furthermore, given the inherent lack of clarity of the *Chandler* approach,[309] it is difficult to anticipate *ex ante* what kind of a relationship would give rise to a duty of care.[310] This, coupled with the fact that the burden of proof in such cases falls on the claimant, means that raising a successful claim proves problematic. For the above reasons, although bypassing limited liability under tort law is an option for some involuntary creditors, it cannot be deemed to be a robust protection mechanism.

(iii) Deterring Companies from Engaging in Hazardous Behaviour - Section 172(1) of the Companies Act 2006

Because limited liability allows companies to easily diversify their risks, it is widely contended that the doctrine incentivises companies to engage in overly hazardous activities.[311] In this respect, it is even argued that, on numerous occasions, limited liability "has caused unlimited harm".[312] In this regard, s.172(1) of the Companies Act 2006 was enacted to encourage a company's board to consider the long-term impact of its decisions, which, in theory, should limit the risk of companies engaging in hazardous activities.

Section 172(1) reflects the inclusive "enlightened shareholder value"[313] approach, which posits that long-term relationships are more beneficial for a company than short-term financial gains.[314] Accordingly, the introduction of

[308] [2014] EWCA Civ 635 at [37] per Tomlinson LJ.

[309] On the contrary, it is sometimes asserted that the lack of clarity of the *Chandler* approach could be ameliorated by the categorisation of the circumstances where the parents' direct tortious liability could be recognised, see Daisuke Ikuta, 'The Legal Measures against the Abuse of Separate Corporate Personality and Limited Liability by Corporate Groups: The Scope of Chandler v Cape Plc and Thompson v. Renwick Group Plc.' (2017) 6 UCL Journal of Law & Jurisprudence 60, 82-88.

[310] Rühmkorf (n 40) 177.

[311] Halpern, Trebilcock, and Turnbull (n 189) 148; Leebron (n 46) 1565; Ribstein (n 28) 81; Muchlinski, 'Limited Liability and Multinational Enterprises: a Case for Reform?' (n 39) 915-916; Lipton, 'The Mythology of Salomon's Case and the Law Dealing With the Tort Liabilities of Corporate Groups: An Historical Perspective' (n 33) 480-481; Ireland, 'Limited Liability, Shareholder Rights and the Problem of Corporate Irresponsibility' (n 49) 838.

[312] Katharina Pistor, 'Limited liability is causing unlimited harm' (*Social Europe*, 11 February 2020) <www.socialeurope.eu/limited-liability-is-causing-unlimited-harm> accessed 28 July 2020.

[313] The Company Law Review Steering Group, Modern Company Law for a Competitive Economy: The Strategic Framework (URN 99/654 DTI 1999) 5.1.11-5.1.12. See also Mary Arden, *Common Law and Modern Society: Keeping Pace with Change* (Oxford University Press 2015) 228; Parker Hood, 'Directors' Duties Under the Companies Act 2006: Clarity or Confusion?' (2013) 13(1) Journal of Corporate Law Studies 1, 16.

[314] Andrew Keay, 'Section 172(1) of the Companies Act 2006: An Interpretation and Assessment' (2007) 28.4 The Company Lawyer 106, 107.

s.172(1) was supposed to bring "a cultural change" in the way in which corporate business was done by placing greater emphasis on long-term considerations and wider stakeholders' interests.[315] Thus, s.172(1) was arguably going to pave the way for "an enlightened"[316] company law system.

On this account, s.172(1) requires a director of a company to "act in the way he considers, in good faith, would be most likely to promote the success of the company for the benefit of its members as a whole, and in doing so have regard, amongst other matters, to" a number of factors.[317] Accordingly, this provision states that the director should no longer act merely for the benefit of the company's owners, but rather he should act in a responsible and forward-looking manner, and should reflect the interests of a wider group of stakeholders in his decision-making.[318] Therefore, s.172(1) aims to discourage management from engaging in hazardous activities, which should subsequently decrease the number of instances where involuntary creditors are harmed by reckless profit-orientated corporate actions.

It is often argued, however, that s.172(1) is ineffective in "any practical sense".[319] For instance, the scope of the director's duty under this section is very vague.[320] Namely, it is ambiguous what 'success of the company' entails and how 'success' ought to be achieved.[321] The focus of the duty is in fact extremely narrow.[322] Moreover, it is not known what the phrase "have regard to" means, and thus it is unclear how specifically directors should "have regard" to the factors listed

[315] See the Ministerial Statement of Rt. Hon Margaret Hodge, Minister of State for Industry and the Regions, UK Department of Trade and Industry (now Department for Business Innovation & Skills), 'Companies Act 2006: Duties of Directors' (Ministerial Statement June 2007) 2. <www.dti.gov.uk/files/file40139.pdf> accessed 28 July 2020. See also Andrew Keay, 'The duty to promote the success of the company: is it fit for purpose?' (2010) University of Leeds School of Law, Centre for Business Law and Practice Working Paper, 11. <ssrn.com/abstract=1662411> accessed 8 August 2020.

[316] Grant (n 45) 255.

[317] See John Birds et al., *Annotated Companies Legislation* (2nd ed, Oxford University Press 2012), para 10.172.02.

[318] Nicholas Grier, 'Enlightened shareholder value: did directors deliver?' (2014) 2 Juridical Review 95, 96.

[319] Elaine Lynch, 'Section 172: a Ground-Breaking Reform of Director's Duties, or the Emperor's New Clothes' (2012) 33(7) The Company Lawyer 196, 202; See also Grant (n 45) 255.

[320] Grier (n 318) 97.

[321] ibid; Andrew Keay, 'Section 172(1) of the Companies Act 2006: An Interpretation and Assessment' (n 314) 109.

[322] Alistair Alcock, 'An Accidental Change to Directors' Duties?' (2009) 30 The Company Lawyer 362, 367.

in s.172(1).[323] Furthermore, the list of factors that should be taken into account by a director is subordinated to the duty to promote the success of the company.[324] As a result, the interests of stakeholders will be given consideration insofar as doing so will ultimately promote the success of the company.[325] Therefore, s.172(1) merely gives "an illusion" that the interests of stakeholders will be considered in the corporate decision-making.[326] In addition, most stakeholders lack *locus standi* to enforce breaches of s.172(1), as only "members of the company", that is, present and future shareholders,[327] can bring a statutory derivative action under Part 11 of the Companies Act 2006.[328] It is therefore unlikely that the shareholders would be willing to enforce breaches of s.172(1) in the name of non-shareholders.

Placing s.172(1) in the context of involuntary creditors' protection, it is difficult for involuntary creditors to rely on this section, as this particular group is not expressly listed under the provision. Even though the list is non-exhaustive, it is unlikely that the company's board would "have regard" to the interests of involuntary creditors, as, in most cases, there would be no economic incentive to do so, and thus giving due consideration to involuntary creditors' interests would not "promote the success of the company". Additionally, involuntary creditors lack standing to enforce the provision, as this possibility is available only to the members of the company. For those reasons, s.172(1) has "raised expectations that it cannot

[323] Andrew Keay, 'Having Regard for Stakeholders in Practising Enlightened Shareholder Value' (2019) 19(1) Oxford University Commonwealth Law Journal 118, 120; Grant (n 45) 255.

[324] Davies and Worthington (n 17) 502-503; Robin Hollington, Hollington on Shareholders' Rights (8th ed, Sweet & Maxwell 2017) para 4-38; Keay, 'Section 172(1) of the Companies Act 2006: An Interpretation and Assessment' (n 314) 108; Hood (n 313) 18-19.

[325] Armour (n 53) 3.72; Talbot (n 108) 141. This issue is well-illustrated by the case of *R. (on the application of People & Planet) v HM Treasury* [2009] EWHC 3020 (Admin) at [34] per Mr Justice Sales.

[326] Georgina Tsagas, 'Section 172 of the Companies Act 2006: Desperate Times Call for Soft Law Measures' in Nina Boeger and Charlotte Villiers (eds), *Shaping the Corporate Landscape: Towards Corporate Reform and Enterprise Diversity* (Bloomsbury Publishing 2018) 146.

[327] See *Gaiman v National Association for Mental Health* [1971] Ch 317 at [330] per Megarry J.

[328] John Lowry, 'The Duty of Loyalty of Company Directors: Bridging the Accountability Gap Through Efficient Disclosure' (2009) 68(3) The Cambridge Law Journal 607, 618; Grier (n 318) 98.

deliver",[329] and thus it cannot be regarded as an effective mechanism which limits the risk of the company engaging in reckless behaviour.

IV. Alternatives to Limited Liability from the Involuntary Creditors' Perspective

The doctrine of limited liability has tremendous economic benefits, as it encourages capital investment in corporate entities by limiting shareholders' personal exposure.[330] The above discussion evidenced, however, that limited liability makes involuntary creditors prone to an increased risk of harm, and that the law does not offer sufficient protection to this particular group of creditors. On this account, it is argued that, in modern times, limited liability is applied beyond its original purpose, as it is no longer merely used as a mechanism that stimulates economic growth.[331] In turn, it allows companies to easily diversify their commercial risks, which exposes involuntary creditors to excessive losses.[332] Consequently, because of the significantly weaker position of involuntary creditors, shifting the risk away from companies onto this specific group seems to be unjustified.[333] In fact, no one would admit that the law on limited liability, created largely in the nineteenth century, appropriately accommodates all social costs involved in the activities of modern multinational corporations.[334] Indeed, in its report on the reform of UK company law, the Company Law Review Steering Group (CLRSG) acknowledged that because a parent company can shield itself from liability for the wrong done by its subsidiary, involuntary creditors are exposed to an increased risk of being left with no compensation.[335] The CLRSG chose, however, not to address this particular problem on the ground that undercapitalisation of subsidiaries is a matter of insolvency law rather than company law.[336] Moreover, it found that there was no evidence that corporate groups abuse the corporate form in order to avoid

[329] Stephen F. Copp, 'S. 172 of the Companies Act 2006 fails people and planet?' (2010) 31(12) The Company Lawyer 406, 408; Philip Ashton, 'How Fred the Shred Got Away with It: Loud Calls for Company Law Reform' (2013) 1 Birkbeck Law Review 187, 200.

[330] Alexander (n 205) 390; Millon (n 30) 1309; Peterson (n 22) 63.

[331] Blumberg (n 12) 575.

[332] Villiers (n 50) 95.

[333] Stefan H.C Lo, 'Liability of Directors as Joint Tortfeasors' (n 18) 121.

[334] Bryan Horrigan, 'Directors' Duties and Liabilities - Where Are We Now and Where Are We Going in the UK, Broader Commonwealth, and Internationally?' (2012) 3(2) International Journal of Business and Social Science 21, 41.

[335] The Company Law Review Steering Group, Completing the Structure (URN 00/1335 DTI 2000) para 10.58.

[336] ibid para 10.59.

liability.[337] It can be argued, however, that the CLRSG's latter conclusion was based on a false assumption; using the corporate structure for the purpose of avoiding liability is sometimes the main reason why the corporate form is implemented.[338] It is therefore worth looking at the alternatives to limited liability, which would reallocate commercial risks back to the company, and could provide involuntary creditors with a better degree of protection in the modern market reality.

With this purpose in mind, it must be noted that the complete abolition of the doctrine of limited liability and the unlimited liability regime resulting therefrom would discourage people from investing their capital in corporate entities,[339] and would severely hamper governance of companies.[340] The unlimited liability system is therefore not desirable from the commercial point of view.[341]

That being said, Leebron acknowledges that limited liability of shareholders seems to be unjustified in relation to tort victims, who are poor risk-bearers.[342] On this account, Leebron,[343] as well as Hansmann and Kraakman,[344] argue that shareholders should be held liable for corporate torts on a pro rata basis. Under this approach, when the tort claim exceeds the assets of the company, the company's shareholders would be held liable for a share of the company's debt proportionate to the share of their equity ownership.[345] Arguably, this approach would improve the position of involuntary creditors, as the company's owners would have an incentive to consider the interests of this particular group of creditors alongside their own. Moreover, the economic benefits of limited liability would not be lost, as this approach would not increase the shareholders' monitoring costs.[346] In fact, as Blumberg observes, a pro rata liability system was in place in California between

337 ibid para 10.59; Rühmkorf (n 40) 183-185. See also Muchlinski, 'Holding Multinationals to Account: Recent Developments in English Litigation and the Company Law Review' (n 207) 173.

338 Rühmkorf (n 40) 183-185; Hansmann and Kraakman (n 32) 1912; Muchlinski, 'Holding Multinationals to Account: Recent Developments in English Litigation and the Company Law Review' (n 207) 168-169.

339 Bainbridge and Henderson (n 21) 59-60; Richard A. Posner, 'The Rights of Creditors of Affiliated Corporations' (1976) 43 University of Chicago Law Review 499, 502.

340 See Stephen M Bainbridge, 'Privately Ordered Participatory Management: An Organizational Failures Analysis' (1998) 23 Delaware Journal of Corporate Law 979, 1055–1057.

341 For a comprehensive critique of the unlimited liability system, see Muchlinski (n 39) 925-926.

342 Leebron (n 46) 1601.

343 ibid 1565-1650.

344 Hansmann and Kraakman (n 32) 1879-1934.

345 ibid 1892-1894. See also Xue Feng, 'Corporate Liability Towards Tort Victims in the Personal Injury Context' (Phd thesis, Queen Mary University of London 2018) 48-53.

346 Leebron (n 46) 1605-1608. For an interesting discussion on the impact of pro rata unlimited liability of shareholders on the market, see Graeme G Acheson, Charles R Hickson, and John D Turner, 'Does Limited Liability Matter? Evidence From Nineteenth-Century British Banking' (2010) 6(2) Review of Law & Economics 247, 247-273.

the nineteenth and twentieth centuries, and the existence of such a system did not hinder the economic growth of this state.[347]

However, the pro rata system has various practical shortcomings. For example, because in certain cases individuals are able to obtain shares for non-cash consideration, imposing financial liability upon them might render the pro rata system highly arbitrary.[348] Moreover, such a system would arguably be ineffective in relation to controlling shareholders, who could dilute their shareholding in the subsidiary, but at the same time they would be able to retain control over the subsidiary through a minority interest.[349] In addition, given the large numbers of shareholders that certain companies have, the *pro rata* system would increase litigation costs for involuntary creditors,[350] and it would be very difficult, if not impossible, for the claimant to enforce pro rata liability of shareholders against foreign corporations for procedural reasons.[351]

On a different note, because involuntary creditors are poor risk-bearers and cannot bargain with the corporation,[352] numerous American scholars argue for this group of creditors to be given preference over the company's secured and unsecured creditors[353] when the company becomes insolvent.[354] According to the proponents of this approach, since voluntary creditors can protect themselves from potential losses more easily,[355] they are much better risk-bearers than involuntary creditors.[356] Additionally, this solution, unlike pro rata unlimited liability, would not suffer from various practical shortcomings, as it only concerns the basic relationship between

[347] Blumberg (n 12) 597-599; Muchlinski (n 39) 926. Similar findings were reached by Patterson, see Mark R Patterson, 'Is Unlimited Liability Really Unattainable: Of Long Arms and Short Sales' (1995) 56 Ohio State Law Journal 815.

[348] Henry G Manne, 'Our Two Corporation Systems: Law and Economics' (1967) 53.2 Virginia Law Review 259, 262.

[349] Muchlinski (n 39) 926.

[350] ibid; Manne (n 348) 262; Nina A Mendelson, 'Control-Based Approach to Shareholder Liability for Corporate Torts' (2002) 102 Columbia Law Review 1203, 1284-88.

[351] Alexander (n 205) 429-431.

[352] French (n 31) 608-609; Leebron (n 46) 1639-40; LoPucki (n 200) 1897-1898.

[353] It is worth noting that in the UK, a company's employees are given preferential treatment on insolvency, see Insolvency Act 1986, Schedule 6 paras 9 and 10.

[354] Leebron (n 46) 1643-1646; LoPucki (n 200) 1913; Barry E Adler, 'A World Without Debt' (1994) 72 Washington University Law Quarterly 811, 825-827; Price (n 46) 470; Rebecca J. Huss, 'Revamping Veil Piercing for All Limited Liability Entities: Forcing the Common Law Doctrine into the Statutory Age' (2001) 70(1) University of Cincinnati Law Review 95, 132-133; Hannoch Dagan, 'Restitution in Bankruptcy: Why All Involuntary Creditors Should be Preferred' (2004) 78 American Bankruptcy Law Journal 247, 277. For a comprehensive discussion of the corporate credit system, see Christopher M.E. Painter, 'Tort Creditor Priority in the Secured Credit System: Asbestos Times, the Worst of Times' (1984) 36 Stanford Law Review 1045, 1045-1085.

[355] Huss (n 354) 133.

[356] Posner (n 339) 503.

the specific groups of creditors.[357] However, giving preference to involuntary creditors on insolvency would mean that various debt-lending institutions, such as banks, would no longer be guaranteed repayment. Consequently, these institutions would likely charge extreme interest rates in order to compensate for the increased risk of non-payment, or would simply not be willing to lend money at all.[358] Such an outcome would have disastrous consequences for companies attempting to raise money through debt.[359] Thus, despite being theoretically sound, this approach would likely not work in practice.[360]

It is also argued that involuntary creditors would be afforded a better degree of protection under a control-based shareholder liability system.[361] Under this approach, joint and several liability would be imposed upon shareholders, who had the 'capacity to control' the company, and can thus be held responsible for the company's acts.[362] The 'capacity to control' would be established in cases, where a given shareholder possesses a large amount of stock in the company, or exercises actual control over that company.[363] Notably, this approach would not suffer from the shortcomings of the pro rata liability system, as it would be able to hold the company's controlling shareholders accountable for the wrongful acts of the company. Moreover, the economic benefits of limited liability would not be lost, as small shareholders, who did not exercise any control over the company, would be exempt from liability.[364]

It is worth noting that this approach would prove particularly useful in the context of corporate groups, which tend to profit the most from externalisation of risks.[365] Namely, according to Muchlinski, a legal presumption of parent responsibility should be introduced to challenge the problems existing under the current law on limited liability.[366] Such a presumption would arise based on the

[357] Price (n 46) 464.

[358] For a discussion of the ways in which secured creditors minimise risks of non-payment and the concept of secured credit in general, see Thomas H Jackson and Anthony T Kronman, 'Secured Financing and Priorities among Creditors' (1979) 88 Yale Law Journal 1143.

[359] This problem is particularly evident in the context of small companies, which prefer to opt against equity financing in order to avoid share dilution.

[360] Such a system has not in fact been implemented anywhere, see Robert K Rasmussen, 'Resolving Transnational Insolvencies Through Private Ordering' (2000) 98(7) Michigan Law Review 2252, 2269.

[361] Mendelson (n 350) 1271-1274. See also Jonathan Crowe, 'Does Control Make a Difference? The Moral Foundations of Shareholder Liability for Corporate Wrongs' (2012) 75(2) The Modern Law Review 159, 159-179.

[362] Mendelson (n 350) 1271-1272.

[363] ibid.

[364] ibid.

[365] Muchlinski (n 39) 918-920.

[366] ibid 923-924.

actual or potential control exercised by the parent company over the subsidiary.[367] Consequently, this control element would give the parent sufficient notice about the potential liability resulting from the acts of its subsidiary.[368] In addition, the onus of proof would be placed on the parent company to rebut the presumption by giving evidence of the independence of the subsidiary, for example, by proving that the third party entered into the transaction with the subsidiary having full knowledge that the parent's liability was limited.[369] This approach would therefore strike a fair balance between companies, their voluntary, and involuntary creditors. Namely, the company would, in appropriate circumstances, have the possibility of rebutting the presumption of liability by proving the independence of the subsidiary; the company's voluntary creditors would be able to assess their commercial risks, and thus they would have an opportunity to take necessary precautions; and the vulnerable involuntary creditors would have a chance of being adequately compensated for their losses, as it would be extremely difficult for the parent company to rebut the presumption of strict liability in relation to this particular group of creditors.[370] Furthermore, because the onus of proof would be on the parent company, this approach would significantly lower litigation costs for involuntary creditors;[371] high litigation costs are in fact one of the biggest shortcomings of the pro rata liability system. Moreover, the presumption of parent liability would enhance the clarity of the law as to the outcome of litigation, which would constitute a significant improvement compared to the current uncertain veil piercing approach.[372] Additionally, this solution would likely incentivise companies to internalise their risks and would consequently deter them from engaging in hazardous activities. Thus, the control-based presumption of parent liability would strike a fair balance between the interests of the various actors involved in the company's activity, and would be able to retain the economic benefits of limited liability.

Therefore, given the considerable benefits of this approach, the unjust outcomes in cases such as *Adams* could be avoided. However, because courts in the United Kingdom are very reluctant to disregard the *Salomon* principle,[373] such a tremendous change in the law would have to be introduced by Parliament.[374] In fact,

[367] ibid 923-924.
[368] ibid.
[369] ibid 924.
[370] ibid.
[371] ibid.
[372] ibid.
[373] Lim (n 280) 483. See, in general, Alan Dignam and Peter B Oh, 'Disregarding the Salomon Principle: An Empirical Analysis, 1885–2014' (2019) 39(1) Oxford Journal of Legal Studies 16, 16-49.
[374] Muchlinski (n 39) 924.

the importance of the *Salomon* case and the doctrine of limited liability cannot be overstated in an evaluation of the development of the UK's capitalist economy. That being said, given the slow transition process of the British economy from a pure profit-orientated system towards a more stakeholder-inclusive one,[375] this change might not be so far away as it seems today.[376]

V. Conclusion

At the end of the previous millennium, the Economist wrote that "the modern world is built on two centuries of industrialisation. Much of that was built by equity finance which is built on limited liability".[377] This is indeed true; limited liability has played a tremendous role in the economic development of the modern world. Namely, by limiting shareholders' personal exposure, limited liability incentivises people to invest in corporate entities and pursue various business endeavours, which in turn stimulates economic development.[378] In this regard, the doctrine is therefore rightly heralded as "the most important characteristic of the modern corporation".[379] The doctrine, however, is also rightly regarded as controversial.[380] Namely, limited liability allows companies to easily externalise their commercial risks, which exposes the vulnerable group of involuntary creditors to significant losses. This problem is particularly evident in the context of corporate groups, where parent companies use the corporate form to insulate themselves from liability for the acts of their subsidiaries. For those reasons, this article has attempted to answer the question of whether involuntary creditors are adequately protected by the current law.

In Part II, this article has analysed the development of the limited liability principle in the UK and has shown that the interests of involuntary creditors were not given adequate consideration at the time of its inception; potentially, the doctrine was never supposed to be applied in relation to this group at all. In Part III, this paper has briefly outlined the economic rationale behind limited liability and has analysed the modern protection mechanisms available to involuntary

[375] Evidenced, for instance, by the introduction of s.172 of the Companies Act 2006.

[376] See Andrew Keay, 'Moving Towards Stakeholderism - Constituency Statutes, Enlightened Shareholder Value, and More: Much Ado About Little' (2011) 22 European Business Law Review 1, wherein it was argued that this transition process cannot be achieved in a blink of an eye, and will likely take time.

[377] 'The Key to Industrial Capitalism: Limited Liability' *The Economist* (London, 23 December 1999) <www.economist.com/finance-and-economics/1999/12/23/the-key-to-industrial-capitalism-limited-liability> accessed 11 February 2021.

[378] Griffin (n 19) 99; Peterson (n 22) 63.

[379] Bainbridge and Henderson (n 21) 19.

[380] Ribstein (n 28) 81.

creditors, such as veil piercing, bypassing limited liability under tort law, and s.172(1) of the Companies Act 2006. It has been found that these mechanisms have numerous shortcomings and do not afford involuntary creditors with an adequate degree of protection. Consequently, Part IV of this paper has evaluated various alternative approaches to limited liability, which could potentially enhance the position of involuntary creditors. Among pro rata liability, giving preference to involuntary creditors on insolvency, and the control-based liability system coupled with the control-based presumption of parent liability, this paper argues for the implementation of the last of these approaches, as it would strike a fair balance between the interests of the various actors involved in a company's activity, and would retain the economic benefits of limited liability. Crucially, given the growing importance of stakeholders' interests in the UK, such a solution might potentially be introduced in the future. On a final note, the limited liability company was a brilliant invention. Similar to various technological novelties, however, once it is applied beyond its purpose, limited liability becomes dangerous.

Cambridge Law Review (2021) Vol VI, Issue i, 182–202

Iraqi Kurdish Self-Determination: A Pathway to Secession? Settling the Questions of Application and Scope

Mohamed Elerian*

Abstract

From the rubble of the U.S. invasion of Iraq, Iraqi Kurds have carved out a degree of de facto political independence that has been largely sheltered from the external interference and ethnic infighting that has plagued Iraq since the fall of Saddam Hussein. This new-found expression of *self-determination* has seen widespread support amongst Iraqi Kurds for the secession of Iraqi Kurdistan and an attainment of a fully-fledged independent Kurdish state. Yet the existence of such a general right to secession does not exist under international law. This paper aims to assess the extent to which Iraqi Kurds are a people with a right to self-determination, and whether that right can express itself through *remedial secession*. It will be submitted that there is insufficient support for the existence of a positive right to remedial secession or for its progressive development under international law, but that even if such a right did exist or were to develop in the future, the situation in Iraqi Kurdistan would not meet the high threshold required for remedial session to be triggered. In light of this, a *political solution* based on a broader autonomy arrangement and increased forms of cooperation is required to resolve the continuing disputes between the Iraqi Federal Government and Iraqi Kurdistan, even if this might not fulfil Iraqi Kurdish demands for statehood. Until

* Intern, United Nations Security Council Practices and Charter Research Branch. LLB (QMUL), LLM (LSE). melerian97@gmail.com.

Iraqi Kurds can rely on regional and external political frameworks that provide the required support for statehood, a Kurdish state will not be viable.

Keywords: self-determination, remedial secession, Iraqi-Kurdistan, political solution, autonomy arrangement

I. Introduction

Despite the growing calamity in Iraq since the 2003 US invasion, the Kurdish region of Northern Iraq ('Iraqi Kurdistan') has enjoyed relative stability and unprecedented levels of self-rule following decades of persecution. Nonetheless, as the Economist describes it,[1] the recent 2017 independence referendum in Iraqi Kurdistan has, however, highlighted how this haven of peace dreams of separating from Iraq's sea of turmoil.

Although in the context of decolonisation, self-determination offered a right to statehood, this has now largely changed as decolonisation has substantially come to an end.

As self-determination has been reformulated into a human right, its form of expression has changed. Self-determination now entails a group of internal rights enjoyed by a people, with the State owing a corresponding duty to protect those rights. Nevertheless, many of those peoples seeking to express their right to self-determination, including Iraqi Kurds, continue to seek to express it through secession and the attainment of independence, even though the existence of such a general right to secession does not exist under international law.[2]

This paper aims to assess the extent to which Iraqi Kurds are a people with a right to self-determination, and whether that right can express itself through remedial secession. Part I of this paper will briefly highlight the reformulation of self-determination into a human right. Part II will then determine whether Iraqi Kurds constitute a people for the purposes of self-determination. After determining that Iraqi Kurds do constitute a people with a right to self-determination, Part III will carry out a brief assessment of the extent to which a right to remedial secession exists under current international law, as well as evaluate the assertion that, even if a right to remedial secession does not currently exist, there is a shift towards its development within international law. It will be submitted that there is insufficient support for the existence of a positive right to remedial secession or for its progressive development under international law, but that even if such

1 'Does Independence Beckon' (*The Economist*, 1 September 2007) <http://www.economist.com/middle-east-and-africa/2007/09/06/does-independence-beckon> accessed 1 May 2020.

2 'How Do You Start a Country?' (*Collectiu Emma*, 5 August 2017) <http://www.collectiuemma.cat/article/2724/how-do-you-start-a-country> accessed 9 May 2020.

a right did exist or were to develop in the future, the situation in Iraqi Kurdistan would not meet the high threshold required for remedial session to be triggered. Through using Iraqi Kurdistan as an example, this paper also seeks to highlight the difficulties with the normative application of remedial secession. The final section (Part IV) will then present an alternative approach to settling the continuing disputes between the federal Iraqi government and Iraqi Kurdistan.

II. Self-Determination as a Human Right

As the legal framework of human rights gained more prominence in international law, self-determination developed as a right applicable to all peoples. In 1966, self-determination was inserted into common Article 1 of the two International Covenants and was to apply to "all peoples". India's reservation that self-determination was to be understood as a right solely of "peoples under foreign domination"[3] was rejected by the Human Rights Committee for violating the treaty's object and purpose.[4] Although Article 1 did little to define the scope of self-determination, the reference to "all peoples" and the fact that the article is found in a human rights treaty intended to have universal applicability strongly suggests a scope beyond that of decolonisation.[5] Regional documents like Article 8 of the Final Act of the Conference on Security and Co-Operation in Europe ('Helsinki Final Act') and Article 20 of the African Charter on Human and Peoples' Rights further underlined the existence of self-determination as a human right applicable to "all peoples". Therefore, in light of the adoption of numerous United Nations ('UN') resolutions, the many States and scholars who have accepted the right of peoples to self-determination,[6] and the fact that most States, including Iraq, have accepted the right to self-determination through their adherence to one or both of the two International Covenants, self-determination exists as a human rights norm of international law applicable to all peoples.[7]

Importantly, international law continues to distinguish between people and minorities; minorities do not have a formal right to self-determination. Therefore,

3 UN Centre for Human Rights, Human Rights: Status of International Instruments (1987) UN Doc ST/HR/5.

4 Report of the Human Rights Committee, 39th Session Supplement No 40 (A/39/40) [142].

5 Hurst Hanum, 'Rethinking Self-Determination' (1993) 34 Virginia Journal of International Law.

6 See, for example James Crawford, *The Creation of States in International Law* (2nd ed., Oxford University Press 2007).

7 Hanum (n 5) 24.

in determining the extent to which Iraqi Kurds have a right to self-determination, it must first be ascertained whether they constitute a people.

III. Determining The 'Self': Iraqi Kurds

Before determining whether Iraqi Kurds constitute a people for the purposes of self-determination, this section will examine the prevalent theories with regards to the identification of a people.

A. The prevailing theories as regards the identification of a people

Although the right to self-determination has continued to develop in international law and has continued to reference a 'peoples' right', there remains no precise definition of the term 'peoples'.[8] Most definitions of the 'self' now include subjective and objective components which emphasise the cultural affinities among a group.[9] At a "minimum",[10] it is necessary for members of the group to think of themselves as a distinct 'people'. Therefore, it is the existence of a collective consciousness which is the subjective factor needed for the people to be identified as a distinct political unit.[11] It is also necessary for the group to have certain objectively-determinable common characteristics. The UNESCO International Meeting of Experts suggested that such characteristics should include a: common historical tradition, racial or ethnic identity, cultural homogeneity, linguistic unity, religious or ideological affinity, territorial connection, and common economic life.[12]

B. Are iraqi kurds 'a people'?

The Kurds are a distinct ethnic group that span across the Middle East's modern borders and who have inhabited the Kurdish mountains since 2000 BC — therefore possessing a clear "territorial connection".[13] The Kurds have manifested in various States, from the Medean Empire in 600BC, to the Ayyubid dynasty, and

8 United Nations Educational, Scientific and Cultural Organization, 'Final Report and Recommendations, The International Meeting of Experts on Further Study of the Concept of the Rights of Peoples' (22 February 1990) UN Doc SHS-89/Conf.602/7 ('UNESCO Report') (para. 21).

9 ibid, para. 23.

10 Hanum (n 5) 57.

11 Thomas D Musgrave, *Self-Determination and National Minorities* (Oxford University Press 2002) [166].

12 UNESCO Report (n 8), para 22.

13 ibid.

the most recent — but short lived — Kurdish Republic of Mahabad in 1946.[14] This distinct historical lineage provides clear and continuous evidence of the Kurdish people as a "distinct ethnicity"[15] since 2000 BC.[16] Even after the formation of the contemporary nation States within the Middle East and the subsequent separation of Kurds from the Ottoman Empire between Iraq, Iran, Syria, and Turkey in the 1920s, Kurds inhabiting Iraq continued to possess distinctive common characteristics. This shared history and continuous distinctive identity despite centuries of upheaval and turmoil represents evidence of an objectively determinable common Iraqi Kurdish ethnicity and history.

This objectively distinct identity of Iraqi Kurds is reflected in their internal and almost homogenous composition. Unlike the wider Iraqi population, "almost all of Iraqi Kurds are Sunni Muslim".[17] In Iraq, 62% of Muslims are Shia,[18] whereas only 2% of Iraqi Kurds are Shia, with 98% being Sunni Muslims.[19] This highlights the clearly distinct and homogenous religious identity of Iraqi Kurds. Additionally, Iraqi Kurds share the common language of Kurdish, which is widely used in the regional administration and education system within Iraqi Kurdistan.[20] Although as a wider language Kurdish does not have a unified script (Perso-Arabic Alphabet and Latinised Alphabet) or dialect (Sorani and Kirmanji), within Iraqi Kurdistan, Sorani is the dominant dialect and a modified Perso-Arabic alphabet is mostly used.[21] Additionally, unlike the rest of Iraq, which predominantly speaks Arabic, within Iraqi Kurdistan, few under twenty-five even understand Arabic,[22] and it has been three decades since Arabic was properly taught in Kurdish schools.[23] This demonstrates a clear unity of language amongst Iraqi Kurds, as well as a linguistic identity that is distinct from the wider Iraqi population.

Iraqi Kurds also satisfy the subjective element of self-determination because they perceive themselves collectively as Iraqi Kurds — a distinct 'people'.[24] The preamble to the Iraqi Kurdish constitution includes terms such as "our people"

14 Akturk, Ahmet Serdar, "Imagining Kurdish Identity in Mandatory Syria: Finding a Nation in Exile" (2013) 866 University of Arkansas.

15 ibid.

16 Alexander Dawoody, 'The Kurdish Quest for Autonomy and Iraq's Statehood' (2006) 41 Journal of Asian and African Studies.

17 Philip S Hadji, 'The Case for Kurdish Statehood in Iraq' (2009) 41 Case W Res J Int'l, 522.

18 Besheer Mohamed, 'Who Are the Iraqi Kurds?' (*Pew Research Center* 2014) <https://www.pewresearch.org/fact-tank/2014/08/20/who-are-the-iraqi-kurds/> accessed 3 May 2021.

19 ibid.

20 Mahir A Aziz, *The Kurds of Iraq* (Bloomsbury Publishing 2011) [165].

21 ibid 78.

22 The Economist (n 1).

23 ibid.

24 Hadji (n 20) 36.

and "nation",[25] thus highlighting the existence of this subjective perception. Since the Gulf War and the enjoyment by Iraqi Kurds of their longest period of self-rule in a century, the common identity of the Kurds has been particularly evident. National symbols are prevalent throughout Iraqi Kurdistan, with Kurdish flags flying throughout the Kurdish region and the Iraqi flag being rarely displayed.[26] In addition, the Kurds have erected statutes and portraits of Kurdish heroes throughout Iraqi Kurdistan.[27] Although these flags and statues are only symbols, they represent tangible indications of a Kurdish sense of common identity that underlines how Iraqi Kurds see themselves as distinct people.

Additionally, Kurds have consistently sought autonomy from the rest of Iraq. Immediately after the fall of Saddam Hussein, the Kurds submitted a proposed Constitution to the Iraqi Governing Council that would give the Kurds the constitutional right to secede from Iraq at any time.[28] Similarly, in the 2017 independence referendum in Iraqi Kurdistan, which had a 72.61% turnout, 92% voted in favour of independence from Iraq. Although the proposal to include a right to secession was rejected by the Iraqi Governing Council and the Iraqi Supreme Federal Court held that no region could secede[29] — thus nullifying the results of the referendum — a majoritarian Iraqi Kurdish desire for autonomy was evident. Iraqi Kurds also have a semi-autonomous region in Iraqi Kurdistan which is run by the Kurdistan Regional Government ('KRG') and recognised as an autonomous region in the Iraqi Constitution.[30] This autonomous political expression indicates that Iraqi Kurds perceive their identity as a people as being distinct, whether in struggling to gain autonomy or in actual autonomy.

Nevertheless, Iraqi Kurds are not entirely cohesive in nature. The strength of tribal and regional factions has often resulted in strong breaks between political parties, with tribal interests overshadowing national ones. The fault line of Kurdish politics runs between the Kurdistan Democratic Party ('KDP') and the Patriotic Union of Kurdistan ('PUK'). Although fighting broke out between both factions in 1994, in the years since, Iraqi Kurdish society has become more united, with a restoration of peace between the two groups. Since 2002, both groups cooperate in the legislative council of Iraqi Kurdistan. While other differences persist, such

[25] Iraqi Kurdistan, The Kurdish Regional Constitution, Apr 19 2004, Preamble.

[26] The Economist (n 1).

[27] Ofra Bengio, *Saddam's Word* (Oxford University Press 2002).

[28] Iraqi Kurdistan, The Kurdish Regional Constitution, Apr 19 2004, Preamble.

[29] Ahmed Rasheed, 'Iraq Court Rules No Region Can Secede After Kurdish Independence Bid' (*Reuters*, 6 November 2017) <https://www.reuters.com/article/us-mideast-crisis-iraq-kurds/iraqcourt-rules-no-region-can-secede-after-kurdish-independence-bid-idUSKBN1D617O> accessed 1 April 2020.

[30] Iraq, The Constitution of Iraq, 15 October 2005, Article 117.

as diverging political ideologies,[31] these differences would not be enough to deprive Iraqi Kurds from constituting a people with legal rights to self-determination. In fact, the rivalry between the KDP and the PUK "has enabled the development of a nascent democratic and pluralistic system",[32] and it is natural that all individuals within a collective do not share the same political preferences.

As a collective unit, Iraqi Kurds have consistently demonstrated a unified preference towards autonomy and a consistently unique and almost homogenous cultural, linguistic, and religious identity. Regardless of passionate divisions between political affiliations, Iraqi Kurds exist as a distinct 'people' within Iraq. Additionally, the collective Iraqi Kurdish desire for independence crosses internal political divisions, as evidenced by the 2017 referendum results. Given that Iraqi Kurds fulfil the objective criteria of a common historical, traditional, ethnic, religious, and linguistic unit, as well the subjective factor of perceiving themselves "collectively as a distinct people",[33] they would constitute a people for the purposes of the legal right to self-determination. Since it has been determined that Iraqi Kurds constitute a people for the purposes of self-determination, the next section will assess whether that right can express itself through remedial secession.

IV. Iraqi Kurdistan and Remedial Secession

As indicated by the 2017 independence referendum, a clear majority of people in Iraqi Kurdistan wish to express their self-determination by seceding from Iraq and forming an independent State. However, the Iraqi Supreme Federal Court has held that no right to secession exists within municipal law, and the federal government has been adamant in refusing to begin talks on a secession agreement with Iraqi Kurdistan. Since an agreement of secession between Iraq and Iraqi Kurdistan does not seem to be an option, the next question is whether Iraqi Kurds — who are a people with a right to self-determination — may express that right under international law by unilaterally seceding from Iraq.

Cassese explains that there are certain defined contexts within which the right to the self-determination of peoples does allow for unilateral secession — namely, it exists for those peoples under colonial rule or foreign occupation.[34] While the right of colonial peoples "to break away from the imperial power is

[31] The KDP tends toward a conservative nationalism, whereas the PUK draws upon a social democratic system.

[32] Ofra Bengio, 'Iraqi Kurds: Hour of Power?' (*Middle East Forum,* Summer 2003) <https://www.meforum.org/554/iraqi-kurds-hour-of-power> accessed 19 May 2020.

[33] Ved P Nanda, 'Self-Determination under International Law: Validity of Claims to Secede' (1981) 13 Case W Res J Int'l L.

[34] Antonio Cassese, *Self-Determination of Peoples* (Cambridge University Press 1998).

now undisputed",[35] Iraqi Kurds are not a people under colonial rule or foreign occupation. Again, although another possible circumstance that might allow for external self-determination involves cases where a State is in dissolution, as in the break of the Former Yugoslavia,[36] Iraq is not in dissolution. The fact that secession may be sanctioned in specific circumstances demonstrates that there is no general prohibition on secession within international law. Nor, on the other hand, is there a general right under international law to unilateral secession. Therefore, the general view is that the law "neither prohibits nor permits unilateral secession outside the specific cases of colonial peoples".[37]

Alternatively, it has been suggested that the right to the self-determination of peoples does allow for unilateral secession in another specific circumstance: through remedial secession. This occurs where a people is "blocked from the meaningful exercise of its right to self-determination internally, it is entitled, as a last resort, to exercise it by secession".[38] Yet, for Iraqi Kurds to possess such a right, there must exist in one of the sources of international law[39] specific provisions providing for a right to remedial secession.

The following section will ascertain whether such a right exists under current international law or whether there is a shift in international law that might suggest that it will progressively develop. This determination will remain brief as, even assuming that there is a right to remedial secession under international law, it will be submitted that the current situation within Iraqi Kurdistan cannot be said to meet the required threshold for remedial secession to apply.

A. Ascertaining the existence of a right to remedial secession under international law

(i) Textual Basis

The textual basis often invoked for a right to remedial secession is the 'saving clause' of the UN General Assembly Declaration on Principles of International Law concerning Friendly Relations and Co-operation among States ('Declaration'). The Declaration now "reflects customary international law"[40] and is therefore a source of binding international law. Under the Declaration a State has to conduct itself "in compliance with the principle of equal rights and self-

[35] Quebec Secession Reference (n 16) [132].
[36] Text of Opinions on Questions Arising from the Dissolution of Yugoslavia (1992) 31 ILM 1494.
[37] Cassese (n 35) 340.
[38] Quebec Secession Reference (n 16) [134].
[39] United Nations, Statute of the International Court of Justice, 18 April 1946, Article 38(1)(b).
[40] *Accordance with International Law of the Unilateral Declaration of Independence in Respect of Kosovo* (Advisory Opinion) [2010] ICJ Rep 403 [80].

determination of peoples"[41] before it is entitled to protection from "any action which would dismember or impair […] [its] territorial integrity or political unity".[42] It is argued that an *a contrario* reading of this clause would suggest that under special circumstances, the principle of self-determination is to be accorded priority over the opposing principle of territorial integrity, thus allowing for remedial secession.

However, even if one were to accept a generous reading of this Declaration which might downgrade a State's right to territorial integrity, Tomuschat explains that any such limitation is far removed from extinguishing a State's territorial integrity by allowing for remedial secession.[43] Therefore, the Declaration is too loosely worded to sanction a positive right to remedial secession.[44]

Alternatively, the Helsinki Final Act does provide a right to "all peoples" to determine their "external political status"[45] within the definition of self-determination. The Act, however, was a non-legally binding regional document that was meant to apply only to the peoples of Europe and therefore cannot be interpreted as being a source of binding international law. In any case, there was no indication that sub-State groups could constitute 'peoples'.[46] Therefore, there is no binding treaty law or non-binding declaration that has developed into a customary norm that establishes a positive right to remedial secession under international law.

(ii) State Practice and Opinio Juris

As regards State practice, Bangladesh's secession is often noted as the classic case of remedial secession. The secessionist movement in Bangladesh was preceded by a brutal governmental policy of repression as well as economic, ethnical, and linguistic discrimination.[47] According to Crawford, the repression carried out by Pakistan's government qualified East Pakistan to be a unit with a right to remedial secession.[48] Tomuschat, however, argues that the secession of Bangladesh was brought about by the principle of effectiveness rather than any legal right to remedial secession.[49] The surrender and withdrawal of the Pakistani Army from East Pakistani territory created a power vacuum that eventually allowed Bangladesh to emerge as a new State on the international stage, particularly with

41 Organization for Security and Co-operation in Europe 'Conference on Security and Cooperation in Europe Final Act' (Helsinki, 1 August 1975) ('Helsinki Final Act') Principle 8.

42 UNGA Res 2625 (XXV) (1970) UN Doc A/RES/2625(XXV) Principle (e) [7].

43 Christian Tomuschat, *Secession and self-determination' in Marcelo G Kohen, Secession: International Law Perspective* (Cambridge University Press 2006).

44 ibid 38.

45 Helsinki Final Act (n 42) Principle 8.

46 Hanum (n 5) 57.

47 Niall Macdermot, 'Crimes Against Humanity in Bangladesh' (1973) 2 The International Lawyer.

48 Crawford (n 5) 142.

49 Tomuschat (n 46) 30.

the political and military support received from India.[50] The argument that the secession of Bangladesh was more a product of the principle of effectiveness than the existence of a right to remedial secession is supported by the silence of the UN on the issue of self-determination,[51] confining itself to demanding that the troops of India and Pakistan be withdrawn from each other's territory.[52] Consequently, Bangladeshi secession cannot serve as an example of unequivocal State support for a positive right to remedial secession, because although Bangladesh remains the only successful case of unilateral secession, its success was more a result of a fait accompli that States (and importantly Pakistan) had to eventually accept.

Alternatively, Kosovo has been argued by some to be an example of remedial secession. Several States proclaimed that Kosovo had a right to remedial secession in their written submissions to the International Court of Justice ('ICJ'), following the UNGA request for an Advisory Opinion on whether Kosovo's declaration of independence was in accordance with international law.[53] Additionally, the Kosovo 'precedent' has been relied upon in the rhetoric of other independence-seeking groups, including South Ossetia and Abkhazia.[54]

However, any State support for Kosovo's right to remedial secession has been limited. The few State submissions to the ICJ supporting Kosovo's right to remedial secession cannot be said to represent the "constant and uniform usage"[55] that is required for the development of customary international law, particularly since eleven States submitted written and oral statements that a right to remedial secession did *not* exist under international law.[56] Even those States that accepted the existence of a right to remedial secession made it clear that Kosovo's situation "was unique and does not set a precedent".[57] Additionally, relying on such submissions as evidence of *opinio juris* is precarious, as they are often based more upon "considerations of convenience or political expediency" than upon general views of the law — which is detrimental to the formation of a customary

[50] Lee C Buchheit, 'Secession. The Legitimacy of Self-Determination' (1981) 14 Verfassung in Recht und Übersee.

[51] ibid 209.

[52] UNSC Res 307 (1971) UN Doc S/RES/307.

[53] Christian Nielsen, 'The Kosovo Precedent And The Rhetorical Deployment Of Former Yugoslav Analogies In The Cases Of Abkhazia And South Ossetia' (2009) 9 Southeast European and Black Sea Studies.

[54] ibid 182.

[55] *Asylum Case (Colombia v Peru)* [1950] ICJ Rep 266.

[56] Nielsen (n 56) 65.

[57] Olli Rehn, 'European Commissioner for Enlargement Introductory remarks on Western Balkans European Parliament, Foreign Affairs Committee' (*European Commission* – 21 March 2007) <https://ec.europa.eu/commission/presscorner/detail/en/SPEECH_07_170> accessed 4 May 2020.

norm.[58] The fact that 103 States[59] have recognized Kosovo's independence should also not be perceived as demonstrating a general recognition of a legal right to remedial secession. Although an analysis of the role of recognition in achieving statehood is outside the scope of this paper, political factors (more so than legal determinations) consistently influence the recognition of States. This is evident by the fact that in the past three years, thirteen countries which had previously recognised Kosovo revoked their recognition following Serbian pressure.[60] In conclusion, the inconsistent State practice and *opinio juris* surrounding Kosovo's unilateral secession, combined with the fact that Kosovo remains to this day an autonomous Republic of Serbia according to the United Nations, underlines the absence of clear-cut international and State practice recognising a positive right to remedial secession.

Other examples where attempts at secession have failed provide ample proof that a right to remedial secession does not exist under international law. For example, in Chechnya the international community universally stressed the need for the preservation of Russia's territorial integrity,[61] even though there is "little doubt that the Chechens qualify as a people"[62] and are alleged to have been the victims of war crimes committed by the Russian forces.[63] As Tomuschat argues, if international law granted a right to remedial secession, the silence of the international community would hardly be understandable in the case of Chechnya.[64]

(iii) Judicial Decisions

Several courts have cited the possible existence of an operative doctrine of remedial secession. These instances date back to the Aaland Islands case, where it was held that the separation of a minority from a State is an "exceptional solution,

58 Dakshinie Gunaratne, 'What Is Opinio Juris – Public International Law' (*Public International law*, April 22 2011) <https://ruwanthikagunaratne.wordpress.com/tag/what-is-opinio-juris/> accessed 19 April 2020.

59 As of April 2020.

60 Craig Turp-Balaz, 'Serbia's Campaign to Reduce the Number Of Countries Which Recognise Kosovo Is Working' (*Emerging Europe*, January 19 2020) <https://emerging-europe.com/news/serbias-campaign-to-reduce-the-number-of-countries-which-recognise-kosovo-is-working/> accessed 19 May 2020.

61 Council of the European Commission 'Declaration on Chechnya' (Strasbourg, 7 October 1999) 0177/1999 [2].

62 David Raič, *Statehood and The Law of Self-Determination* (Kluwer Law International 2002).

63 Amnesty International, 'Brief Summary of Concerns About Human Rights Violations in the Chechen Republic', 1 April 1996, EUR/46/20/96, available at <https://www.refworld.org/docid/3ae6a9c52c.html> accessed 09 May 2020.

64 Tomuschat (n 46) 37.

a last resort" when the "State lacks either the will or the power to enact and apply just and effective guarantees".[65] This seems to provide tacit judicial support for a right to remedial secession. Additionally, Meller argues that the ICJ in the Kosovo Advisory Opinion "unintentionally developed a new concept of the doctrine of remedial secession"[66] by finding that there is no explicit prohibition of universal declarations of independence under general international law.[67] However, it is far from clear how that corresponds to developing a right to remedial secession, particularly since the Court explicitly chose to only note that remedial secession was subject to "radically different views"[68] without pronouncing on the doctrine any further.

In regional and national cases, the Canadian Supreme Court identified subjugation in the non-colonial context as a possible third ground for secession.[69] However, the Court noted that it "remains unclear" whether the right to remedial secession "reflects an established international law standard".[70] Meanwhile, The African Commission on Human and Peoples' Rights ('ACHPR') seemed to suggest that remedial secession existed as a test under international law, and that had there been evidence of "violations of human rights to the point that the territorial integrity of Zaire should be called to question", a peoples would be permitted exercise a variant of self-determination that is "not compatible with the sovereignty and territorial integrity of Zaire".[71] This 'variant of self-determination' could take the form of remedial secession. A similar conclusion was reached in *Kevin Mgwanga Gunme et al v Cameroon*.[72] However, these cases relied upon the African Charter and

[65] The Aaland Island Question: Report Submitted to the Council of the League of Nations by the Commission of Rapporteurs, League Doc. B. 7. 21/68/106 (1921) 4.

[66] Samuel Ethan Meller, 'The Kosovo Case: An Argument for a Remedial Declaration of Independence' (2012) Ga J Int'l & Comp.

[67] *Advisory Opinion on Kosovo* (n 43) [438].

[68] ibid [83].

[69] Quebec Secession Reference (n 16) [133].

[70] ibid [135].

[71] *Katangese Peoples' Congress v Zaire* (1995) ACmHPR Comm 75/92.

[72] *Kevin Mgwanga Gunme et al v Cameroon ('Mgwanga')* Communication 266/03 (2009) [170], [179].

the court gave no indication that its reasoning was influenced by international norms.

(iv) The Progressional Development of a Right to Remedial Secession

Although remedial secession is cited, particularly by courts, in light of the absence of its positive formulation in any of the sources of international law, it cannot be said to exist as a positive right under international law.

At the same time, however, there is no complete affirmation within the sources of international law specifically *prohibiting* a right to remedial secession. This has led some to suggest that although a positive right to remedial session does not currently exist under international law, the development of doctrines such as the responsibility to protect ('R2P') and humanitarian intervention have shifted international law towards basing a State's territorial integrity on its ability to comply with its obligation to protect its people's internal rights. This therefore speaks in favour of the ongoing development of a right to remedial secession under international law.[73]

An in-depth analysis of R2P or humanitarian intervention is beyond the scope of this paper. Nevertheless, such a shift in international law is far from clear. While the enforcement and development of R2P, particularly in light of the 2011 UNSC-authorised military intervention in Libya (which referenced R2P for the first time),[74] might speak in favour of a shift in international law, the application of R2P should be differentiated from both humanitarian intervention and remedial secession. R2P can ground its application in the UNSC's Chapter VII powers, but humanitarian intervention and remedial secession lack a strict legal basis allowing for their application under international law. Therefore, even if R2P were to be applied more consistently, that development would not necessarily reflect a movement in international law that would allow for doctrines like humanitarian intervention or remedial secession to develop by analogy. This is because, unlike R2P, they are not applied through the UNSC's Chapter VII powers and at the current time lack the widespread support needed to develop as stand-alone rights under international law. Even if one did not make that differentiation between the doctrines, "R2P is a political doctrine and not a formal or even material source of international law"[75] and does not seem likely to apply more consistently following its application in Libya, where NATO allegedly overstepped its mandate to effect regime change.[76] This has resulted in a growing distrust of the R2P doctrine by

[73] See e.g., Summers James, 'Relativizing Sovereignty: Remedial Secession and Humanitarian Intervention in International Law' (2010) 6(1) St Antony's International Review.

[74] UNSC Resolution 1973 (2011) UN Doc S/RES/1973.

[75] Chris O'Meara, 'Should International Law Recognize a Right of Humanitarian Intervention?' (2017) 66 International and Comparative Law Quarterly.

[76] ibid 464.

the international community.[77] With regards to humanitarian intervention, its status is far from clear in current international law, as the Non-Aligned Movement (representing the majority of States), as well as China and Russia, have continuously registered strong opposition to its existence within international law.[78] Such opposition is only countered by limited regional and national support,[79] which is not enough to override the UN Charter's prohibition on the use of force in Article 2(4) or the principle of non-intervention enshrined in Article 2(7). Therefore, this does not suggest a shift in international law that may allow for a right to remedial secession to develop.

Other developments focused on the protection of human rights law, including in international humanitarian law and international criminal law, accord with the UN Charter's general emphasis on the peaceful settlement of disputes between nations. Therefore, their development does not signal towards a parallel development of a right to remedial secession or humanitarian intervention, which both call into question "cardinal principles of international law"[80] — namely, the obligation to respect the territorial integrity of States. This is particularly so when one considers the sensitivity of such doctrines to the manipulative influence of States, as was evident when NATO overstepped its mandate in Libya,[81] or when Russia claimed a right to humanitarian intervention for its activities in Ukraine.[82] Additionally, opposition by States containing populations with aspirations of secession (e.g., Spain, Iraq, Morocco) would prevent the right of remedial secession from developing in treaty law or crystallising into a customary norm.

Although the right to remedial secession therefore cannot be said to exist or be progressively developing under international law, the next section will assess whether Iraqi Kurds would be able to trigger a right to remedial secession if the doctrine were to develop under international law.

B. Would iraqi kurds have a right to remedial secession?

Although international law remains unclear on the strict requirements needed for a people's to exercise a right to remedial secession, it is a right that

77 ibid 464.

78 ibid 470.

79 For example, the recognition by some NATO states of the right to humanitarian intervention following the bombings in Kosovo.

80 United Nations Press Office, Press Release (24 April 1991) SG/SM/4560.

81 O'Meara (n76) 464.

82 Transcript of President Putin's interview, Moscow (*Washington Post*, March 4 2014) <https:// https://www.washingtonpost.com/world/transcript-putin-defends-russian-intervention-in-ukraine/2014/03/04/9cadcd1a-a3a9-11e3-a5fa-55f0c77bf39c_story.html/> accessed 19 May 2020.

seems to exist only as a "last resort"[83] when the parent State is unable or unwilling to "enact and apply just and effective guarantees".[84] Furthermore, it only arises when a people are 'blocked' from the meaningful exercise of their right to self-determination internally[85] — including, for example, through the "large-scale and persistent violations of basic human rights".[86]

For a period of 30 years, between 1961–1991, Iraqi Kurds faced extreme persecution and were deprived of their right to "freely determine their political status and freely pursue their economic, social and cultural development".[87] During the Anfal campaign, "Iraqi armed forces […] systematically destroyed more than four thousand Kurdish villages",[88] used chemical weapons and organized the execution of Kurdish civilians.[89] As many as 182,000 Iraqi Kurds were killed, with Human Rights Watch declaring that "Iraq's crimes against the Kurds amount to genocide' with an 'intent to destroy, in whole or in part".[90] This was combined with the destruction of the rural Kurdish economy and infrastructure.[91] Therefore, Iraqi Kurds were deprived of their internal rights to self-determination as Iraq failed in its corresponding obligation under international law to protect those rights.

However, the establishment of the KRG in 1991, as well as the fall of Saddam Hussain's regime in 2003, has brought about a period of unprecedented Iraqi-Kurdish autonomy and socio-economic and political freedoms. Today, Iraqi Kurds have enjoyed their longest period of self-rule in a century, with clear assurances in the Iraqi constitution of a right to participation in the federal government and a right to internal autonomy within Iraqi Kurdistan.[92] Additionally, the federal government is conducting itself in compliance with the principle "of equal rights and self-determination of peoples without distinction".[93] When comparing Iraqi Kurdish autonomy with other agreements proposing special status of greater autonomy, such as the Act on the Autonomy of Åland and the Good Friday Agreement, it is evident that Iraqi Kurdistan enjoys much of the same functional

83 Quebec Secession Reference (n 16) [134].
84 Aaland Island, Rapporteurs (n 66) 4.
85 Quebec Secession Reference (n 16) [134].
86 Allen E Buchanan, *Justice, Legitimacy, And Self-Determination* (Oxford Political Theory 2003).
87 International Covenant on Civil and Political Rights, 23 March 1976, 999 UNTS 171 ('ICCPR') Article 1.
88 George Black, Human Rights Watch, 'Genocide in Iraq: The Anfal Campaign Against The Kurds' (Human Rights Watch 1993) <https://www.hrw.org/reports/1993/iraqanfal/ANFAL-PRE.htm> accessed 7 May 2020.
89 ibid 3.
90 ibid 87.
91 ibid 6.
92 Iraq, The Constitution of Iraq, 15 October 2005, Preamble, Article 4, Article 117, Article 141.
93 Quebec Secession Reference (n 16) [136].

sovereignty.[94] Like Northern Ireland and the Åland Islands, Iraqi Kurdistan enjoys a democratic legislative assembly, as well as executive and independent control over internal administration. Iraqi Kurdistan also enjoys a proactive, if not independent, foreign policy, with 29 countries having a diplomatic presence in the Kurdistan region.[95]

At the federal level, the PUK and the KDP — the two main Iraqi Kurdish political parties — currently hold 43 seats in the Iraqi Parliament and have consistently been elected to the Iraqi Council of Representatives since 2003. At present, the President of Iraq, the Finance Minister, the Housing & Reconstruction Minister, and a member of the Iraqi Federal Supreme Court are all Kurds. Therefore, Iraqi Kurds freely determine their political status as they occupy prominent positions within the federal government and are equitably represented in legislative, executive, and judicial institutions. This political autonomy is also coupled with socio-economic independence, which has led Iraqi Kurdistan to enjoy "more stability, economic development, and political pluralism than the rest of the country".[96] Iraqi Kurds freely make political choices and pursue economic, social, and cultural development.

Continuing issues between Iraqi Kurdistan and the federal Iraqi government, such as limitations in the supply of armaments from the federal government to the Peshmerga forces (the military forces of Iraqi Kurdistan)[97] or Kurdish not being equally treated as an official language in Baghdad, are not enough to give rise to a right to remedial secession. These issues seem to be more of a critique of a federal government lacking resources following decades of turmoil, than of a federal government that is denying Iraqi Kurds their internal right to self-determination.[98]

The right to remedial secession will only develop as a matter of last resort if Iraq "lacks either the will or the power to enact and apply just and effective guarantees".[99] While Iraqi Kurds might have been able to rely on such a right of last resort during the Saddam regime, this threshold cannot be currently met in light of the guarantees of internal self-determination given to Iraqi Kurds. Just as the Canadian Supreme Court held that the 'exceptional circumstances' needed for a right to remedial secession "are manifestly inapplicable to Quebec under

94 Hanum (n 5) 66.

95 'Department of Foreign Relations, Kurdistan Regional Government' (*Kurdistan Regional Government*, 2020) <https://gov.krd/dfr-en/> accessed 3 April 2020.

96 Kawa Hassan, 'Kurdistan's Politicised Society Confronts A Sultanistic System' (August 2015) Carnegie Endowment for International Peace.

97 Maria Fantappie, 'The Peshmerga Regression' (Crisis Group, 14 June 2015) <https://www.crisisgroup.org/middle-east-north-africa/gulf-and-arabian-peninsula/iraq/peshmerga-regression> accessed 10 May 2020.

98 The Economist (n 1).

99 Aaland Island, Rapporteurs (n 66) 4.

existing conditions",[100] these same exceptional circumstances are also 'manifestly inapplicable' to Iraqi Kurdistan under existing conditions. Accordingly, even though Iraqi Kurds are a people, Iraqi Kurdistan and its representative institutions would not possess a right under international law to secede unilaterally from Iraq.

C. A critique of remedial secession

As mentioned above, a right to remedial secession appears to be a right that exists as a last resort once a State's behaviour has caused "an unbridgeable gap for finding realistic and effective alternatives to remedial secession".[101] Therefore, there is, inadvertently, a requirement of strict temporal proximity between the triggering of a right to remedial secession by a people on the one hand, and the appalling violations of a people's right to self-determination on the other. This requirement of 'last resort', however, provides only a superficial solution because it seems to be disproportionately dependent on the short-term situation of a people without taking into account a holistic image containing past persecutions and long-term denial of a right to self-determination. Rather paradoxically, a people would probably only be able to effectuate a right to remedial secession once they are organised enough to speak with one voice. Without foreign military assistance (as in Bangladesh), this level of organisation cannot be formed during times of persecution and denial of a people's political and social rights. Yet, by the time a group has gained those rights adequately enough to effectively demand remedial secession, a right to remedial secession would have 'run out', as the home State would be conducting itself in compliance with the principle of equal rights and self-determination of peoples. This also signals to States who have enacted long-term policies of socio-economic and political discrimination against a people that a short-term remedy of increased rights would prevent any right to remedial secession from arising.

Nevertheless, a right to remedial secession must be a strict right of last resort that only arises once a State has proved either unable or unwilling to enact effective guarantees of internal rights to a people. Now that decolonisation has substantially come to a conclusion, self-determination is rightly more frequently expressed through autonomy arrangements rather than secession as populations today are far more riddled with competing claims. A wide right to remedial secession would raise the prospect of endless carving-out as new minority groups emerge. For example, if Iraqi Kurdistan were to secede, this might give rise to similar demands of secession from the Yazidi population living predominantly within Iraqi Kurdistan. Therefore, there are legitimate concerns between making secession more broadly available on the one hand and increased fragmentation and instability on the other. What this demonstrates is that a clarification of the cogent rules and principles

[100] Quebec Secession Reference (n 16) [138].
[101] Rai (n 63) 78.

of remedial secession or its crystallisation into an international law norm would not change the fact that remedial secession seems impossible to apply because there are multiple competing interests at stake (including, in this case, those of the Iraqi federal government, Iraqi Kurds, Yazidis, and regional powers) that cannot be reconciled through the absolute approach of remedial secession. Additionally, an attempt at clarifying the legal scope of the law on remedial secession on the international stage runs the risk of limiting or completely prohibiting unilateral secession in the name of international peace and security.

Instead of attempting to clarify the legal or normative scope of remedial secession, there should be a clarification of the current law on self-determination by first setting out a precise definition of a 'people' as well as advancing an approach towards settling claims of self-determination that is primarily focused on protecting internal rights, increasing minority rights and resolving claims of secession through autonomy arrangements and negotiated solutions. Additionally, the distinction between internal and external self-determination should be scrapped, as these two modes of achieving self-determination are inherently inter-connected (e.g., a violation of internal rights to self-determination would give rise to demands for external self-determination).[102] By scrapping this distinction, there would be a greater emphasis on ensuring the protection of self-determination as a human right — which international law does provide for — instead of an excessive focus on external self-determination, which international law largely does not regulate. It is the approach of negotiated solution mentioned above that the next section will consider as the answer to settling the ongoing disputes between the Iraqi federal government and Iraqi Kurdistan.

V. The Future of Iraqi Kurdistan

The further realisation of the right of Iraqi Kurds to self-determination should take the form of increased cooperation with the Iraqi federal government and a broader autonomy arrangement. To understand the form that such a broader autonomy arrangement would take, it is important to first determine the problems that persist between Iraqi Kurdistan and the federal government — namely, those surrounding oil and disputed internal Kurdish-Iraqi boundaries. The Iraqi Constitution stipulates that Baghdad must give 17% of national oil revenues to Iraqi Kurdistan, but the KRG argues that this provision excludes newly-discovered oil fields, over which it claims full control. As for the territorial dispute, the KRG claims an area that exceeds its official boundaries, particularly surrounding the

[102] Otto Kimminich, 'A "Federal' Right of Self-Determination?" in Christian Tomuschat (ed), *Modern Law of Self-Determination* (M. Nijhoff 1993).

oil-rich Kirkuk region. Importantly, even if Iraqi Kurds had a right to remedial secession, it is not clear how independence would solve the persisting problems with Iraq. For example, disputes persist between South Sudan and Sudan regarding the division of oil revenues and the disputed region of Abyei[103] even though South Sudan has seceded.

Instead, Iraqi Kurdish demands for further self-determination could be realised through institutional intergovernmental bodies between Iraqi Kurdistan and Iraq. These would take the form of regular and frequent meetings between the federal government and the KRG to promote cooperation at all levels of government, as well as to seek negotiated solutions to the current disputes. Additionally, on matters officially not devolved to Iraqi Kurdistan (such as foreign policy), intergovernmental bodies would allow Iraqi Kurdistan to put forward proposals. An Iraqi-Kurdish Council made up of ministerial representatives from the Iraqi federal government and the KRG could also be established to promote cooperation and the creation of common policies, much in the same way as the British-Irish Council. Such high-level cooperation would result in an interlocking and interdependence between Iraq and Iraqi-Kurdistan, ensuring that the success of each depends on that of the other. This would also protect Iraqi Kurdistan and Iraq from the further reification of ethnic differences. On 3 May 2020, following economic difficulties caused by COVID-19 and low oil prices, a delegation of Kurdish officials travelled to Baghdad with the aim of "strengthening Erbil-Baghdad" ties for the first time in over a year.[104] It is similar but more consistent and structured cooperation that will allow Iraqi Kurdish demands for further self-determination to be met.[105]

It is doubtful, however, whether anything short of independence would fulfil the Iraqi Kurdish visceral inclinations towards statehood, as evidenced by the consistent polls showing significant support towards Iraqi Kurdish independence regardless of increased autonomous rule. An unconstitutional declaration of

[103] 'South Sudan Profile' *(BBC News*, 6 August 2018) <https://www.bbc.com/news/world-africa-14069082> accessed 2 May 2020.

[104] Viktor Katona, 'Iraqi Kurdistan On The Brink Of Collapse As Oil Prices Crash' (*OilPrice.com*, 3 May 2020) <https://oilprice-com.cdn.ampproject.org/v/s/oilprice.com/Energy/Energy-General/Iraqi-Kurdistan-On-The-Brink-Of-Collapse-As-Oil-Prices-Crash.amp.html?usqp=mq331AQFKAGwASA%3D&amp_js_v=0.1#referrer=https%3A%2F%2Fwww.google.com&amp_tf=From%20%251%24s&ampshare=https%3A%2F%2Foilprice.com%2FEnergy%2FEnergy-General%2FIraqi-Kurdistan-On-The-Brink-Of-Collapse-As-Oil-Prices-Crash.html> accessed 10 May 2020.

[105] Milena Sterio, 'Self-Determination and Secession Under International Law: The Cases Of Kurdistan And Catalonia' (*Asil.org*, 5 January 2018) <https://www.asil.org/insights/volume/22/issue/1/self-determination-and-secession-under-international-law-cases-kurdistan> accessed 11 May 2020.

secession leading to *de facto* secession for Iraqi Kurdistan, while potentially fulfilling this inclination in the short-term, would not be tolerated by the international community even though "general international law contains no applicable prohibition on declarations of independence".[106] This is clear from the response of the international community to the 2017 independence referendum. The States with the closest bilateral trade relations with Iraqi Kurdistan — Turkey and Iran — denounced the referendum, with Turkey calling it a "terrible mistake"[107] and Iran labelling it a "Zionist plot".[108] The wider international community also made clear that the dispute between Baghdad and Erbil must be resolved by finding a formula of "coexistence within the Iraqi State".[109] This international hostility is driven by a fear that Iraqi Kurdish statehood would result in Kurds in neighbouring States striving towards creating a greater Kurdistan which would "disturb international peace and security".[110] Therefore, as long as Iraq is unwilling to authorise a Kurdish independence referendum and to negotiate a separation agreement, secession for Iraqi Kurdistan will not become a reality. Increased cooperation with the federal government is the way forward.

VI. Conclusion

The extreme persecution faced by Iraqi Kurds, combined with a denial of internal rights of self-determination that has spanned several decades, has led some to mistakenly assert following the 2017 independence referendum that, by virtue of their right to self-determination, Iraqi Kurds would have a right to secede under international law.[111] However, this paper has demonstrated that a people with a right to self-determination (as the Iraqi Kurds are) do not, under current international law, have a positive right to remedial secession, and that it does not seem that international law is moving in a direction that would allow for such a right to develop. But even if such a right to remedial secession were to exist under

[106] *Advisory Opinion on Kosovo* (n 43) [84].

[107] 'Ankara Ve Bağdat'tan IKBY Referandumuna Tepki' (*DW.COM*, 9 June 2017) <https://www.dw.com/tr/ankara-ve-ba%C4%9Fdattan-ikby-referandumuna-tepki/a-39176559> accessed 7 May 2020.

[108] 'Barzani Middleman For Zionists To Partition Islamic Countries: Velayati' (*Press TV*, 26 September 2017) <https://www.presstv.com/Detail/2017/09/26/536571/Iran-Iraq-Kurdistan-Ali-Akbar-Velayati-Massoud-Barzani-Zionists-KRG-independence-referendum> accessed 2 May 2020.

[109] Anna Borshchevskaya, 'In Search Of A New Patron, The KRG Turns Back To Moscow' (*The Washington Institute*, 1 June 2018) <https://www.washingtoninstitute.org/policy-analysis/view/in-search-of-a-new-patron-the-krg-turns-back-to-moscow> accessed 2 April 2020.

[110] ibid.

[111] See e.g., Gregory J Ewald, 'Kurd's Right to Secede under International Law: Self-Determination Prevails over Political Manipulation' (1994) 22 Denv J Int'l Law & Pol'y.

international law, Iraqi Kurds would not meet the threshold required to trigger remedial secession.

Additionally, by using Iraqi Kurdistan as an example, the fundamental issues with the normative and legal scope of the application of remedial secession have been demonstrated. It is submitted that remedial secession is not and should not develop as a right under international law. Instead, there should be an increased focus on viewing self-determination as a human right and resolving secession claims through autonomy arrangements. By virtue of that, it is a broader autonomy arrangement and increased forms of cooperation that will resolve the continuing conflict between Iraq and Iraqi Kurdistan, even if this might not fulfil Iraqi Kurdish demands for statehood.

Finally, just because secession is not viable at the moment, this does not mean that an Iraqi Kurdish State will not be viable in the future. Frames of reference regarding statehood have consistently changed over time. We have gone from 40 States at the founding of the UN to 193 now, with the existence of nation-states such as Belgium or Luxembourg seeming impossible in the 19th century. Therefore, the viability of an Iraqi-Kurdish State is not a question that can be answered absolutely. If Iraqi Kurdistan can rely on regional and external frameworks in the future, much in the same way that Luxembourg and Belgium rely on NATO and the EU, then Iraqi Kurdistan might possess the regional support required to effectuate statehood.[112]

[112] Collectiu Emma (n 2).

Cambridge Law Review (2021) Vol VI, Issue i, 203–235

Marking the Internal and External Limits of Discrimination Law in *Lee v Ashers Baking Company*

Emily M L Ho*

Abstract

One of the most frequently occurring clashes between different groups in society is where religious beliefs concerning homosexuality are manifested in the public square through positive acts such as preaching against homosexual practices and omissions such as a refusal to provide goods and service to homosexual individuals. In cases such as these, discrimination law is expected to intervene to uphold the value of equality. *Lee v Ashers Baking Company* was no different, involving bakers who refused to fulfil a customer's order of a cake iced with the message 'Support Gay Marriage'. The Supreme Court decided in favour of the bakers, and in so doing, analysed and marked the limits of discrimination law — specifically, the prohibition of direct discrimination. This article seeks to mark these limits, examining their desirability against the background of domestic and international jurisprudence and political theory concerning freedoms of religion and expression. It first examines the *internal limits* of discrimination law, namely the different fact patterns in which the conventional 'shape' of direct discrimination cases has been permitted to be modified. It then examines the *external limits* of discrimination

* BA (Cantab), LLM candidate (Harvard), eho@llm22.law.harvard.edu. I am grateful to Professor Trevor Allan for supervising this paper. All errors remain my own.

law, namely the pressure exerted on the reach of discrimination law by alleged discriminators' freedoms of religion and expression.

Keywords: discrimination law, direct discrimination, LGBTQ discrimination, freedom of religion, freedom of expression

I. Introduction

As the equality project advances, diversity in the United Kingdom increases, and political polarisation becomes starker, clashes become increasingly frequent between groups of different race, belief, gender, and sexual orientation. In these instances, equality law is expected to intervene. One of the most paradigmatic clashes is where religious beliefs concerning homosexuality are manifested in the public square through positive acts such as preaching against homosexual practices,[1] and omissions such as a refusal to provide goods and services to homosexual individuals.[2] *Lee v Ashers Baking Company*[3] embodies the latter clash. Ashers Bakery, a business run according to its owners' — the McArthurs'— Christian beliefs, cancelled an order placed by Mr Lee for a cake iced with the message "Support Gay Marriage". It did so because, in the owners' view, fulfilling the order would be promoting a message that was contrary to their beliefs, violating their conscience.

The type of clash embodied in *Lee* is particularly challenging because it presses at the limits of discrimination law, from both the inside and the outside. The internal limits are faced because *Lee* presents a unique pattern of alleged discrimination: the differential treatment was dealt out irrespective of the specific customer's identity, thereby bending the conventional form of direct discrimination as differential treatment of *persons*. To have found discrimination in *Lee* would therefore have expanded the range of conduct prohibited. The external limits are faced because Ashers' unilateral objection to the order was rooted in their religious belief that marriage is reserved for heterosexual couples, engaging their freedoms of religion and expression. The aim of this article is to utilise this special *Lee* fact pattern to trace the limits of discrimination prohibitions. These should then be heeded to maintain the conceptual integrity of discrimination law, and vindicate the values of a liberal plural society in future discrimination cases.

This article focuses on the impact of the decision in *Lee* on the application of the Equality Act 2010 in England and Wales, whose provisions are substantially

1 *Hammond v DPP* [2004] EWHC 69 (Admin), [2004] 1 WLUK 95.
2 *Bull v Hall* [2013] UKSC 73, [2013] 1 WLR 3741.
3 *Lee v Ashers Baking Company* [2018] UKSC 49, [2018] 3 WLR 1294.

analogous to the Northern Ireland provisions applied in *Lee*.[4] In the first Part, I draw the internal limits of discrimination law by analysing and evaluating the Supreme Court's application of the tools of comparator, indissociability to protected characteristics, associative discrimination, and indirect discrimination. There, I conclude that the limits of the discrimination concept observed by the Court can be explained as a sustained focus on the personal characteristics of individuals, rather than the substance of messages involved, even when the substance relates to protected personal characteristics. In the second Part, building on the Court's brief analysis of relevant rights in the European Convention on Human Rights, I study the tension between discrimination law and freedoms of religion and of expression, concluding that the Court rightly observed these external limits. I supplement the Court's brief reasoning with an analysis of case law in other jurisdictions and propose a preferable future trajectory for the interaction between these values for future discrimination cases.[5]

II. Internal Limits

A. Link between protected characteristiscs and less favourable treatment

The Supreme Court's analysis of direct discrimination firmly refocuses discrimination prohibitions as protections against differential treatment of *persons*, and clarifies what that means. This was captured by Lady Hale's terse statement in her judgment that "[by] definition, direct discrimination is treating people differently".[6] This seems a rather obvious point until one confronts the dispute at the heart of *Lee*, which raises questions about what it means to discriminate against a person. Lee claimed he had been treated less favourably on grounds of his sexual orientation or political beliefs by being refused his order, whilst Ashers claimed they had not treated Lee less favourably on those grounds but rather objected to the message requested regardless of Lee's personal characteristics, and would have so objected whatever the customer's characteristics. The Supreme Court decided in favour of Ashers on both the grounds of sexual orientation and political belief,

[4] Fair Employment and Treatment (Northern Ireland) Order 1998 (SI 1998/3162 (NI 21)); Equality Act (Sexual Orientation) Regulations (Northern Ireland) 2006 (SI 2006/439).

[5] At the time of writing, it has been reported that the European Court of Human Rights will soon hear a claim by Mr Lee on the implications of the UK Supreme Court ruling on his rights under the European Convention on Human Rights. The court will pronounce on whether the UK has fulfilled its obligations to protect Mr Lee's Convention rights. I explore these issues in Section III.

[6] *Lee* (n 3) [23].

drawing a distinction between discriminating against a person and discriminating against a message:

> "[i]n a nutshell, the objection was to the message and not to any particular person or persons [...].[7] There was no less favourable treatment on [the ground of political beliefs] because anyone else would have been treated in the same way. The objection was not to Mr Lee because he, or anyone with whom he associated, held a political opinion supporting gay marriage. The objection was to being required to promote the message on the cake. The less favourable treatment was afforded to the message not to the man".[8]

It is a reasonable first impression of the decision that the distinction is artificial and formalistic. Is it not the point of discrimination law to foster an environment in which diversity is tolerated and even celebrated, so that any individual can obtain goods and services and take part in society without hindrances like the one faced by Lee in this case? It is argued, however, that the distinction can be supported.

The deconstructed issue faced by the Court was: what is the requisite link between the protected characteristic in question and the less favourable treatment afforded? This is an essential aspect of any direct discrimination claim because not all "less favourable treatment" is simply characterised as unlawfully discriminatory: the treatment must be specifically related to a prohibited ground (i.e., protected characteristic).[9] Campbell and Smith have called this the "grounding requirement"[10] for direct discrimination. The question of what this requisite link is can only be answered satisfactorily with reference to the aims of discrimination and equality law, which is the focus of this section. More analytically, the Court's granular application of the relevant tools of discrimination law — discussed in later sections — are best understood in light of its circumscription of this requisite link, thereby internally limiting discrimination law.

The standard link between protected characteristic and treatment afforded in core direct discrimination cases is that of less favourable treatment dealt out simply because the specific recipient of the treatment is of a particular race, sex, age, or other protected characteristic. Take the example of a plumber who only fixes

7 ibid [34].

8 ibid [47].

9 The protected characteristics in the Equality Act 2010 (at section 4) are the characteristics possession of which have been determined as illegitimate grounds for treating individuals differently.

10 Colin Campbell and Dale Smith, 'The Grounding Requirement for Direct Discrimination' (2020) 136 LQR 258.

the plumbing in white households, and refuses to carry out a service requested by an Indian household. The Indian household has received less favourable treatment simply because they are not white: they have been differentiated expressly on grounds of their race (a protected characteristic) and dealt with accordingly. This standard link is not the only possible permutation that gives rise to a claim in direct discrimination. An example of an expansion of the range of possible links is associative discrimination, where the protected characteristic that has factored into the less favourable treatment belongs to a person other than the recipient of the treatment, but who is associated with that recipient. Another example is perception-based discrimination, which finds unlawful discrimination where the recipient of the treatment does not actually possess a protected characteristic, but where the alleged discriminator thought they did possess it, and discriminated on that ground. These expansions show how the requirement of a characteristic-treatment link has not been rigidly interpreted.

Some links are impermissible, however, not simply because they are tenuous but because they distort the very concept of discrimination. For example, counsel for Lee at the County Court submitted that "under [the Fair Employment and Treatment Order], discrimination can take place on the grounds of the discriminator's religious belief and political opinion".[11] Thus, counsel claimed that the protected characteristic could belong to the alleged discriminator who dealt out the less favourable treatment. In rejecting this proposal, Lady Hale laid out the basic requirements for the characteristic-treatment link:

> "[t]he purpose of discrimination law is to protect a person (or a person or persons with whom he is associated) who has a protected characteristic from being treated less favourably because of that characteristic. The purpose is not to protect people without such a characteristic from being treated les [*sic.*] favourably because of the protected characteristic of the alleged discriminator […].[12]
>
> [Such] a reading would be inconsistent with article 3(2)(a) [the provision which prohibits direct discrimination] which requires a comparison between the person receiving the less favourable treatment and 'other persons': this would not be possible if the treatment were on the grounds of the discriminator's beliefs

[11] *Lee* (n 3) [42]; Lee *v Ashers Baking Co Ltd* [2015] NICty 2, [2015] 5 WLUK 483 [47(7)].
[12] *Lee* (n 3) [43].

because everyone would be treated alike".[13]

Lady Hale's rejection of this proposal is based on an understanding of the meaning of discrimination as contravention of the rule that 'everyone [should] be treated alike'. This is a notably *personal* understanding of discrimination: it is not established where there has been less favourable treatment which was related — in the alleged discriminator's mind — to an unfavourable perception of a race, sex, sexual orientation, or other characteristic, conceived in the abstract. It is only established where a protected characteristic is possessed (or perceived to be possessed) by a specific individual who is either themselves less favourably treated, or is associated with someone who is.

The reason for requiring a characteristic-treatment link lies in the rationale underpinning discrimination law. The idea of direct discrimination as a wrong is based on the principle of formal equality that 'like cases be treated alike'. This Aristotelian principle of consistent treatment has underpinned the notion of equality from its inception.[14] There have been various moral justifications proposed for this principle, one of the most significant of which is that of universal human dignity, deriving from Aquinian philosophy. This theory, based on dignity, holds that inconsistent treatment fails to respect the universal human dignity of individuals by refusing to confer advantages on them on the basis of characteristics that are irrelevant, especially where the characteristic has been the subject of historical prejudice.[15]

Dignity continues to be regarded as the basis of discrimination law. The Universal Declaration of Human Rights ties equality — a foundational ideal of the Declaration — to dignity: "All human beings are born free and equal in dignity and rights".[16] The EU Charter of Fundamental Rights also emphatically declares the value of dignity: "Human dignity is inviolable. It must be respected and protected".[17] Although the specific content of the concept of 'dignity' is not clear, it is agreed that the basic idea is "recognition of the worth of the human person as a fundamental principle".[18] This "worth of the human person" is upheld

[13] ibid [44].

[14] Jarlath Clifford, 'Equality' in Dinah Shelton (ed.), *The Oxford Handbook of International Human Rights Law* (OUP 2013) 420, 422.

[15] Patrick Shin, 'Is There a Unitary Concept of Discrimination?' in Deborah Hellman and Sophie Moreau, *Philosophical Foundations of Discrimination Law* (OUP 2013) 163.

[16] Universal Declaration of Human Rights, Article 1.

[17] EU Charter of Fundamental Rights, Article 1.

[18] Christopher McCrudden, 'Human Dignity and Judicial Interpretation of Human Rights' [2008] European Journal of International Law 655, 710, quoting Paolo Carozza, 'My Friend is a Stranger: The Death Penalty and the Global Ius Commune of Human Rights' (2003) 81 Texas Law Review 1031, 1081.

by discrimination law, which prohibits the singling out of individuals from others solely because they have a particular characteristic.

This universal dignity is not harmed where everyone is treated alike. The contrast between *Lee v Ashers* and *Bull v Hall* is illustrative here. In *Bull v Hall*, Christian B&B hoteliers operated a policy of only allowing married couples to book a double room, on the basis of their religious belief that sexual relations should only take place within marriage. Therefore, they refused to allow the complainants, a gay couple, to book a double room. At the time, there was no legislative provision for same-sex marriage. This differed from *Lee* in a crucial respect. In *Lee*, any customer ordering that cake would have been refused, and so everyone would have been treated alike. The personal worth of the customer as a human being sharing universal dignity equally with any other human being, would have been respected. The bakery would have made no distinction between customers who made that order, their decision to refuse being based solely on the content of the message. Indeed, if the order had come in through a nameless online form, it would still have been refused. On the contrary, in *Bull*, a heterosexual couple would have been able to book the double room, whereas the homosexual couple (complainants) were not allowed to. This is clearly differential treatment of individuals personally, because the hoteliers' conduct differed according to the personal characteristics of the customer in question. Such differential, non-universal treatment does not respect the universal dignity of the heterosexual and homosexual customers alike, because the former have been regarded as being entitled to the benefits of a double room, but not the latter. The attitude of the hoteliers — whether malicious or wholly well-meaning — is immaterial.[19] It may be argued that *Bull* is not all that different from *Lee* because both instances of conduct sprang from a religious belief applied as a blanket policy for all their business operations. However, the very reason that *Lee* is a unique matrix is not that Ashers held a protected belief, but that it was absolutely immaterial what the sexual orientation or political belief of the customer before them was: the order was refused not because Lee was gay or because he supported same-sex marriage, but because that message had been ordered. In contrast, in *Bull* the sexual orientation of the customers was absolutely material: it determined whether the benefits of the double room would be conferred or not.

The Canadian judicial explication of basic dignity helpfully sets out its connection to the principle of consistent treatment. Courts have regarded a violation of dignity to have taken place where people have been treated differently — on a personal and specific level — on the grounds of a protected characteristic,

[19] *R(E) v JFS Governing Body* [2009] UKSC 15, [2010] 2 AC 728.

thereby relegating them to a demeaned position in the community of individuals they live in:

> "[e]quality means that our society cannot tolerate legislative distinctions that treat certain people as second class citizens, that demean them, that treat them as less capable for no good reason, or that otherwise offend fundamental human dignity".[20]

In *Bull*, the homosexual couple were effectively relegated as 'second class citizens' relative to the hypothetical heterosexual couple that would occupy the 'first class' in this metaphor, because on grounds of their sexual orientation they could not obtain the benefit they sought. In contrast, in *Lee*, no customer was first or second class: they were all treated alike. The capacity of the equal marriage message itself to perpetuate first-class and second-class citizenship is beyond the scope of this inquiry. Firstly, that operates on a secondary plane of analysis (since it does not concern consistent treatment of customers) and is therefore irrelevant. Secondly, it is also an inappropriate factor for judicial analysis since the inquiry into the dignity aspect of equal marriage rights entails a more politically debatable question of dignity compared to whether treatment of individuals has been consistent. This is especially so considering that same-sex marriage had not yet been legalised in Northern Ireland at the time. The level of basic dignity being protected by the legal discrimination concept is that of being viewed on the same level as other individuals despite possession of a protected characteristic.

A historical appraisal of discrimination law supports this view of prohibited discrimination. Fredman has observed that the concept of equality in the UK gained traction "with the advent of mercantile capitalism and the loosening bonds of feudalism", and the significance of the equality principle in that period was its economic outworking in "the principle of freedom of contract" or the "notion of equal parties".[21] This exposes the most basic understanding of the dignity of human beings, which requires that they not be viewed as occupying different levels of society simply by virtue of arbitrary differences, but rather as equal, and treated accordingly in transactions and social interactions. In Shin's words, this basic equality principle prohibits adverse treatment on the basis of "an antagonistic attitude toward individuals because of a [protected characteristic]".[22]

In short, principle and history inform us that discrimination prohibitions have the narrow aim of addressing violations of the basic universal dignity of human beings that requires consistent treatment. This should not be carelessly

[20] *Law v Canada* [1999] 1 SCR 497, [51].
[21] Sandra Fredman, *Discrimination Law* (2nd edn, OUP 2011) 5.
[22] Shin (n 15) 173.

disregarded in favour of expansive interpretations of discrimination law, as this risks impinging on the autonomy of individuals — specifically their freedoms of expression and religion. The nobility of the ideal of equality, and the openness and indeterminacy of the general social concept of discrimination that is used to describe a range of undesirable behaviours in society, can lead an enthusiastic judge to engage in a teleologically expansive interpretation of discrimination law, widening the scope of situations that could be found to constitute unlawful discrimination. It is commendable that the judges did not so expand the law in *Lee*.

This principled narrow domain of discrimination law requires a continued tethering of the relevant protected characteristic to the recipient of the less favourable treatment (or their associate). The granular tools of discrimination law with which counsel for Lee attempted to draw a qualifying characteristic-treatment link, and the Supreme Court's application of them, merit detailed discussion as they collectively operate as the internal limits of discrimination law.

B. Choice of comparator

The choice of comparator — an essential step in claims of direct discrimination — is often not obvious, as demonstrated in the different choices made initially by Brownlie J in the County Court[23] and then by the Northern Ireland Court of Appeal,[24] affirmed by the Supreme Court.[25] The latter, in rejecting the former's formulation of the comparative exercise, clarified that the relevant protected characteristic had to be possessed by a specific individual: either the one who was meted out the less favourable treatment or an associated individual. No other circumstance in the comparator counterfactual can change, for otherwise direct discrimination may be established on facts that actually lack a qualifying link between the adverse treatment and the protected characteristic.

The competing comparator options in *Lee* yielded dramatically different outcomes. In the sexual orientation claim, Brownlie J compared Lee to a "heterosexual person placing an order for a cake with the graphics either 'Support Marriage' or 'Support Heterosexual Marriage'".[26] This was rejected by Morgan LCJ in the Court of Appeal because it "changed both the sexual orientation of the person and the message".[27] Instead, "[the] true comparator was a heterosexual person seeking the same cake". The Supreme Court upheld this.[28] This appellate

[23] *Lee v Ashers Baking Co Ltd* [2015] NICty 2, [2015] 5 WLUK 483 [42].
[24] L*ee v Ashers Baking Co Ltd* [2016] NICA 39, 2016 WL 06268003 [24].
[25] *Lee* (n 3) [24], [34]–[35].
[26] *Lee* (n 23) [42].
[27] *Lee* (n 24) [24].
[28] *Lee* (n 3) [24], [47].

conclusion circumscribed the appropriate comparator to account only for the personal characteristics of individuals, without changing any other circumstance. In the Equality Act 2010, this circumscription is expressly mandated by section 23(1): "[o]n a comparison of cases for the purposes of section 13 […] there must be no material difference between the circumstances relating to each case". Even without express provision, this approach accords with the principle underpinning direct discrimination law. The mischief to be remedied, after all, is denial of equal dignity to an individual because of their protected characteristic.

This was carefully and commendably applied in *Ladele v Islington LBC*.[29] Ms Ladele, a registrar for the council, refused to register civil partnerships on the grounds of her religious belief. Whilst the employment tribunal had adopted the comparator of a gay registrar (who would have registered the partnerships) and thereby found direct discrimination, the Employment Appeal Tribunal (EAT) and the Court of Appeal held that the tribunal had adopted the wrong comparator. Instead, the appropriate comparator is "another registrar who refused to conduct civil partnership work because of antipathy to the concept of same sex relationships but which antipathy was not connected [to] or based upon […] her religious belief".[30] If Ladele was treated differently from this comparator, then it would be clear that she had been treated differently, contrary to discrimination law, because of her religious belief. If the tribunal's comparator had been adopted, differential treatment would as plausibly be attributed to her religious belief as to her bare non-compliance with her employer's instruction (which is a legitimate basis for discipline).

The District Judge's comparator in *Lee* suffered from a similar fault as the tribunal's improper comparator in *Ladele*. If Ashers' policy was to refuse a gay customer's "Support Gay Marriage" order and to fulfil a heterosexual customer's 'Support Heterosexual Marriage' order, there are two possible reasons for such differential treatment, one directly discriminatory and the other not: Ashers could be said to have treated Lee as it did either because Lee was personally gay, or because they did not want to print the message. The latter has nothing to do with Lee's sexual orientation or his identity generally — "anyone else would have been treated in the same way"[31] — albeit it triggers the discussion (in the second Part below) of Ashers' freedoms to do so. To become certain that it was discriminatory conduct, the comparator would need to hold the message constant and vary only Lee's sexual orientation, hence the Supreme Court's choice of comparator. Setting up the comparative apparatus as the Supreme Court did would secure the required

[29] *Ladele v The London Borough of Islington* [2009] EWCA Civ 1357, [2010] 1 WLR 955.
[30] ibid [39].
[31] *Lee* (n 3) [47].

link between the treatment and the protected characteristic. It would ensure that direct discrimination is not established simply because a less favourably treated complainant happens to possess a certain protected characteristic. As Lord Nicholls warned in *Nagarajan v London Regional Transport*,[32] the "crucial question" is "why the complainant received less favourable treatment": it could have been on grounds of race, or because he was not as qualified for the job; it could have been on grounds of religious belief, or because she did not comply with general instructions; it could have been on grounds of sexual orientation, or because the message ordered was unfavourable to the bakers.

C. Indissociability

One avenue of relaxing the requirement of a characteristic-treatment link is the recognition that the alleged discriminator need not have overtly expressed their criterion for treatment to have been the protected characteristic. If their overt criterion was indissociably linked to a protected characteristic, such that the criterion was simply a proxy for differential treatment based on that characteristic, they have directly discriminated. This get-around has been helpfully explained by Benn as an "anti-formalistic device",[33] to ensure that treatment which is genuinely discriminatory in substance will not be found to be non-discriminatory simply because the discriminator has framed their criterion in terms that do not — on their face — refer to a group with a protected characteristic.

In *Lee*, Lady Hale held that the criterion determining Ashers' treatment of Lee — i.e., that he had ordered a cake with the message "Support Gay Marriage" — was not indissociable from being gay (sexual orientation), but that it might have been indissociable from his support of equal marriage rights (political belief). Nevertheless, a reading of the statute that is compatible with Convention rights had to be adopted, and so it was regarded as not being so. Therefore, Lee could not rely on the doctrine of indissociability to construct the requisite link between his protected characteristic and the adverse treatment he had received. It is argued that this was a correct decision. A significant critique that has been levelled against it has been its apparent inconsistency with the application of indissociability in *Bull v Hall*, the most recent significant case on the doctrine at the time, where the gay couple's marital status was held to be indissociable from their sexual orientation given that same-sex marriage was not, at the time, legal.[34] However, a closer analysis

[32] *Nagarajan v London Regional Transport* [2000] 1 AC 501, 511.

[33] Alex Benn, 'The UK Supreme Court and the Gay Marriage Cake: Is "Indissociability" Half-baked?' (*OxHRH Blog*, January 2019), <https://ohrh.law.ox.ac.uk/the-uk-supreme-court-and-the-gay-marriage-cake-is-indissociability-half-baked/> accessed 22 February 2021.

[34] ibid.

in this section of the facts in each case will reveal that Lady Hale's reasoning was consistent. If Lady Hale had found indissociability in *Lee* (as in *Bull*), that would have constituted an unprincipled extension of *Bull*. It would have deformed the concept of direct discrimination by overly liberalising the characteristic-treatment link, and in so doing, enmeshed it with indirect discrimination.

A central case of indissociability is presented in *James v Eastleigh Borough Council*.[35] The council applied a policy of allowing free entry into the public swimming pool for those over the statutory retirement age, which was, at the time, 60 for women and 65 for men. This meant that men between 60 and 64 years of age could not gain free entry into the pool whereas women in that age range could. Although, as Lady Hale stated, "the criterion used for allowing free entry […] was not sex but statutory retirement age", the criterion had the effect in substance of affording different treatment to men and women. The subset of individuals between 60 and 64 who are beyond their retirement age exactly corresponds with the subset representing women, and conversely, the subset of individuals who have not yet met their retirement age exactly corresponds with the subset representing men. This means that users of the swimming pool were treated differently by virtue of their protected characteristic of sex. It did not matter that the council had no discriminatory intent.[36] The effect of the policy was the same as if they had overtly stated that, for those between 60 and 64 years of age, women could enter for free whereas men had to pay a fee (which would more obviously constitute direct discrimination).

The doctrine of indissociability is susceptible to incremental extensions when applied to criteria that are increasingly dissociable from a protected characteristic. Unchecked, these extensions may carry into unprincipled findings of direct discrimination. In *Eastleigh BC*, the criterion was logically indissociable from — or a direct proxy for — the characteristic of sex. It was impossible, as a simple matter of categories, for a 60 to 64-year-old man to have passed his statutory retirement age. In contrast, in *Bull*, the overt criterion was not as logically dissociable from the protected characteristic in question. The overt criterion applied by Mr and Mrs Bull, Christian B&B hoteliers, was that only married couples could book double accommodation as a matter of their preference.[37] At the time, homosexual couples could not be married. Lady Hale held that the criterion of having to be married was indissociable from being of heterosexual orientation, and adversely treated homosexual individuals on the basis of sexual orientation. She acknowledged "that some people of homosexual orientation can and do get married, while […] some

[35] *James v Eastleigh Borough Council* [1990] 2 AC 751.
[36] See now *JFS* (n 19).
[37] *Bull* (n 2) [9].

people of heterosexual orientation can and do enter civil partnerships",[38] but held that this fact could be "[left] aside" because marriage and civil partnership[39] may be regarded as analogous legal institutions for the flourishing of heterosexual and homosexual relationships respectively, and therefore "the criterion of marriage or civil partnership [may be regarded] as indissociable from the sexual orientation of those who qualify to enter it". The criterion and characteristic in *Bull* are at least a shade less indissociable than those in *Eastleigh BC*, and required the — not necessarily tenable — assumption that those in heterosexual marriages are of a heterosexual orientation, in addition to the family policy considerations.[40]

Failing to limit extensions such as those in *Bull* risks internally distorting the concept of direct discrimination. When applied traditionally as in *Eastleigh BC*, the doctrine of indissociability maintains an acceptable link between the adverse treatment and protected characteristic. It prohibits policies that effectively stratify society, adversely treating whole swaths of individuals personally possessing a protected characteristic, even when the overt policy does not stratify as such. (Although the link may not have been subjectively drawn in the alleged discriminator's mind, it is firmly established that good intentions and motives do not vindicate direct discriminators.[41]) The doctrine therefore remedies the mischief addressed by the direct discrimination prohibition, i.e., differential treatment that falls precisely on lines of certain characteristics. This meets the overarching aim of upholding equal dignity and status, as maintained in the previous section. Particularly where equality has been compromised by historical circumstances and prejudices, to effectively purge these "antecedent inequalities"[42] it is necessary to counter differential treatment on these lines even if the alleged discriminator's motives were absolutely benign.

Where the criterion and characteristic are more dissociable, however, to establish direct discrimination on grounds of indissociability would be expansive. Such an application of the doctrine may blur the distinction between direct and indirect discrimination. Granted, the distinction is already partially eroded by the very existence of this doctrine of indissociability; however, it cannot be allowed to do so any more than is necessary to vindicate the rationale explained above. In a discrimination case in the European Court of Justice (ECJ), *Schnorbus v Land*

[38] ibid [29].

[39] At the time, marriage was lawful only for heterosexual couples and civil partnership was lawful only for homosexual couples.

[40] *Bull* (n 2) [26]–[29].

[41] *JFS* (n 19).

[42] Fredman (n 21) 13.

Hessen,[43] Advocate General Jacobs distinguished between direct discrimination established by means of the indissociability doctrine and indirect discrimination:

> "[t]he discrimination is direct where the difference in treatment is based on a criterion which is either explicitly that of sex or necessarily linked to a characteristic indissociable from sex. It is indirect where some other criterion is applied but a substantially higher proportion of one sex than of the other is in fact affected".[44]

The blurring of the boundary in *Bull* is clear when Advocate General Jacobs' distinction is applied to its facts. Lady Hale's recognition that individuals of homosexual orientation do enter heterosexual marriages seems to fit better with a finding of indirect discrimination: a "substantially higher proportion" of heterosexually married individuals are of heterosexual orientation, and a "substantially higher proportion" of homosexual couples were not married since this was not legally possible. It is not as tenable to regard sexual orientation as "necessarily linked" to whether or not one is in a legal marriage.

Perhaps this explains the apparent reluctance of the ECJ in its jurisprudence to find direct discrimination by means of the doctrine, preferring instead to find indirect discrimination. In *Schnorbus*, applicants for practical training to be employed in the civil service in Hesse, Germany were given priority if they had completed compulsory military or civilian service. However, German law only required men to complete compulsory military service. Advocate General Jacobs advised that the criterion of completing military service was not indissociable from being female because the relationship between the criterion and the characteristic was attributed to a legislated policy and not to an unchanging fact of nature such as the relationship between pregnancy and being female.[45] This point was memorably expressed in these terms: "No amount of legislation can render men capable of bearing children, whereas legislation might readily remove any distinction between men and women in relation to compulsory national service".[46] Therefore the preference in favour of national service was not "as such" a preference in favour of men over women. The Court followed Advocate General Jacobs' analysis and found indirect discrimination instead. Ten years later in *Bressol v Gouvernement de la Communauté Française*,[47] Advocate General Sharpston advised that the criterion of having a right of residence in Belgium was indissociable from being of Belgian

[43] Case C-79/99 *Schnorbus v Land Hessen* [2000] ECR I-10997.
[44] ibid [33].
[45] ibid [40].
[46] ibid [40].
[47] Case C-73/08 *Bressol v Gouvernement de la Communauté Française* [2010] 3 CMLR 559.

nationality since Belgian nationals acquired that right automatically whilst non-nationals had to meet additional requirements to do so. Therefore, she advised, the policy directly discriminated on grounds of nationality.[48] The Court declined to follow this, and found instead that the policy was indirectly discriminatory.

The significance of the direct/indirect distinction is that direct discrimination cannot—whereas indirect discrimination can — be objectively justified by showing that the provision, criterion, or practice is "a proportionate means of achieving a legitimate aim".[49] Addressing both direct and indirect discrimination furthers the aim of equality. Direct discrimination focuses more narrowly on formal equality, or equality of treatment.[50] Indirect discrimination focuses on a more substantive notion of equality, aiming at equality of opportunities or of outcomes for different groups in society. As Fredman writes: whereas direct discrimination focuses on equal treatment, indirect discrimination "recognises that equal treatment may itself have a disparate impact"; therefore "it is the disparate impact of an apparently neutral requirement that establishes a prima facie case of indirect discrimination".[51] Given these differences, it is clear that direct discrimination is the "more overt form of discrimination".[52] It is adverse treatment of a group of individuals filtered by their protected characteristic, such that one may draw a Venn diagram representing groups with different characteristics and find that the differential treatment follows those groups exactly. This is more harmful than indirect discrimination, where, owing to factors which may or may not be ascertainable,[53] a policy has particularly disadvantaged a group defined by a protected characteristic, but the outcome for individuals in that group was not *as such* linked to their protected characteristic.[54]

It is fair that central cases of indissociability such as *Eastleigh BC* should be construed as direct discrimination. As explained above, the policy manifested the evil of stratified treatment of individuals. Furthermore, if the policy had been held to be indirectly discriminatory, the council might have objectively justified the

[48] *Schnorbus* (n 44) [67]–[68].

[49] Equality Act 2010, section 19(2)(d).

[50] Bob Hepple, *Equality: The New Legal Framework* (Hart Publishing 2011) 54.

[51] Sandra Fredman, 'The Reason Why: Unravelling Indirect Discrimination' (2016) 45 ILJ 231.

[52] Jane Mair, 'Direct Discrimination: Limited by Definition?' (2009) 10 International Journal of Discrimination Law 3, 13.

[53] *Essop v Home Office* [2017] UKSC 27, [2017] 1 WLR 1343.

[54] The structural distinction of indirect discrimination from direct discrimination—the treatment not needing to be linked as such to protected characteristics—can be seen especially in the Supreme Court's landmark decision in *Essop* (n 53) that a claimant does not need to be able to explain how a provision, criterion, or policy led to disparity. Therefore, on the facts, a career promotion assessment that was failed by a significant proportion of non-white candidates could be found to be indirectly discriminatory without the claimant having to show which particular aspects of the assessment disadvantaged non-white candidates.

policy according to the aim of helping pensioners and the differential treatment would not have been remedied, simply because the direct discrimination did not appear on the face of the policy. It is in cases such as this that direct discrimination — the more overt, unjustifiable species of discrimination — should be established: where whole swaths of individuals set apart by a protected characteristic are filtered out by the discriminator's criterion and treated differently. It is only in such cases that indissociability can be permissibly applied. It is argued that *Bull* represents the weakest acceptable indissociable link between the discriminator's criterion and the protected characteristic, and therefore the outer boundary of the doctrine of indissociability. It is the exceptionally weighty policy reasons that justifies the outcome in *Bull*. That a legal union is a personal act fairly taken as a manifestation of one's sexual orientation is an assumption that undergirds marriage policy; further, civil partnership at the time was regarded as the institution analogous to marriage for homosexual couples.

On the contrary, the criterion applied in *Lee* was not sufficiently indissociable: the criterion of ordering the message "Support Gay Marriage" is not proxy-linked to the person's sexual orientation or political belief. There is no similar policy reason why someone who orders a cake with a custom message should be regarded as necessarily advocating the message personally (i.e., possessing that protected political belief), let alone as belonging to the group for which the message expresses favour (i.e., possessing a homosexual orientation). Regarding political beliefs, perhaps Lee's personal assistant, who does not hold that belief, was sent to order the cake on his behalf;[55] or perhaps Lee is ordering the cake as a gift for his neighbour's QueerSpace party, but does not support gay marriage himself. These are conceivable situations that make it clear that Ashers' criterion did not as such exclude a whole swath of individuals with a certain political belief for different treatment. The criterion more readily approximates indirect discrimination, which is discussed below.

Even more conceivable are situations where the person who orders the cake is not homosexual themselves. Lady Hale herself highlighted: "People of all sexual orientations, gay, straight or bi-sexual, can and do support gay marriage. Support for gay marriage is not a proxy for any particular sexual orientation".[56] The recent legalisation of same-sex marriage in Northern Ireland, Hambler has observed, is vivid evidence that "support [for gay marriage] went rather wider than simply the

[55] It is acknowledged, however, that this might raise an issue of agency.

[56] *Lee* (n 3) [25].

gay community alone".[57] Whilst the Court of Appeal did remark that "[there] was an exact correspondence between those of the particular sexual orientation and those in respect of whom the message supported the right to marry",[58] it is unclear how this link is significant for the purposes of establishing a direct discrimination claim. On the contrary, for what it is worth, to limit the legitimate supporters of equal marriage rights to those who are themselves gay would do no favours to the LGBTQ movement. In sum, given the dissociability of sexual orientation and of political belief from ordering the message, the Supreme Court correctly held that direct discrimination could not be made out.

In light of the danger of deforming direct discrimination by expansively applying the doctrine, it is argued that the Court should not even have entertained the possibility of indissociability of the message from political belief.[59] As has been shown, there are strong reasons internal to discrimination law why the criterion was not indissociable from the protected characteristic, and so the Court need not have made its call based only on their section 3 Human Rights Act 1998 duty to uphold Convention-compliant interpretations of the law. To have established indissociability — which presumably it would have done if Convention rights happened not to have been engaged — would have foregone the requirement of an acceptably close characteristic-treatment link, finding discrimination beyond the principled internal limits of direct discrimination.

D. Associative discrimination

Associative discrimination is another tool for establishing a characteristic-treatment link, where adverse treatment is dealt out because of a protected characteristic belonging not to the individual who has received the treatment but to individual(s) associated with them. The Supreme Court's application of this doctrine in *Lee* was another valuable internal delimitation of the concept of direct discrimination.

The Court of Appeal had held that even if Ashers did not perceive that Lee was gay, they had discriminated because he was perceived as associating with "the gay and bisexual community". As the less favourable treatment was dealt out because of the sexual orientation of that community, it was held that this was associative direct discrimination.[60] The Supreme Court disagreed. Firstly,

[57] Andrew Hambler, 'Cake, Compelled Speech, and a Modest Step Forward for Religious Liberty: the Supreme Court Decision in Lee v Ashers' (2018) 181 Law & Justice – Christian Law Review 156 (in relation to a statistic on the nearby Irish referendum).

[58] *Lee* (n 24) [58].

[59] *Lee* (n 3) [48].

[60] *Lee* (n 24) [58].

there been "no evidence that the bakery had discriminated on that or any other prohibited ground in the past". On the contrary, there was evidence that Ashers "employed and served gay people and treated them in a non-discriminatory way" in the course of their business. Therefore, there was insufficient factual basis for inferring that Ashers had discriminated on this ground against Lee's associates. What was far clearer was that "[the] reason [for their conduct] was their religious objection to gay marriage".[61] Secondly, there needed to be a "closer connection" than simply that "the reason for the less favourable treatment has something to do with the sexual orientation of some people".[62] Lady Hale expressly refrained from defining the closeness of the association required to find associative discrimination. It is argued that, in future cases, this connection should be narrowly construed.

The classic example of associative discrimination is presented by *Coleman v Attridge Law*[63] where the claimant, who formerly worked as a secretary for a law firm, alleged that she had been "subject to unfair constructive dismissal and had been treated less favourably than other employees because she was the primary carer of a disabled child".[64] The ECJ held that the principle of equal treatment in Directive 2000/78 applied not only to individuals who themselves have a disability. Direct discrimination had taken place because the claimant was treated less favourably "based on the disability of [her] child, whose care is provided primarily by [her]".[65] She had received adverse treatment because of a protected characteristic belonging to an individual associated with her.

The ambiguity that remained after *Coleman* and other associative discrimination cases (and left unresolved after *Lee*) is: what connection or association must there be between the person possessing the protected characteristic and the treatment afforded? Or — deconstructed — what constitutes a sufficient characteristic-treatment link?

A wide view of associative discrimination regards the requisite connection as between the treatment afforded and *a protected characteristic in the abstract*. The important connection to establish is between the reason for the treatment and any protected characteristic, where it is immaterial who possesses the protected characteristic and how they are linked to the treatment. The emphasis is not on the association between the person with the characteristic and the recipient of the treatment, but simply on the existence of a hypothetical group of people possessing a protected characteristic, who are discriminated against by the alleged

61 *Lee* (n 3) [28].
62 ibid [33].
63 C-303/06, *S. Coleman v Attridge Law and Steve Law* [2008] ECR I-05603; [2007] IRLR 88.
64 ibid [22].
65 ibid [56].

discriminator's conduct. On one reading of the case, the ECJ in *Coleman* upheld this wide view:

> "[The] purpose of [Directive 2000/78], as regards employment and occupation, is to combat all forms of discrimination on grounds of disability. The principle of equal treatment enshrined in the directive in that area applies *not to a particular category of person* but *by reference to the grounds* mentioned in Article 1 (emphasis added)".[66]

More starkly, Advocate General Maduro opined:

> "[The] Directive performs an exclusionary function: it excludes religious belief, age, disability and sexual orientation from the range of permissible reasons an employer may legitimately rely upon in order to treat one employee less favourably than another. In other words […] it is no longer permissible for these considerations to figure in the employer's reasoning when she decides to treat an employee less favourably".[67]

These statements locate the centre of gravity of associative discrimination in the characteristic in the abstract, rather than the individual possessing that characteristic.

The ECJ further widened this view in *CHEZ Razpredelenie Bulgaria AD v Komisia za zashtita ot diskriminatsia*,[68] showing how far the flexibility of the wide view can expand the scope of discrimination law. CHEZ, an electricity supplier in Bulgaria, generally installed electricity meters 2-metres high except in one Roma-majority district, where the meters were 6- to 7-metres high. The reason for the distinction was to prevent electricity theft by tampering with the meters which, CHEZ argued, occurred more frequently in that district. The complainant, a non-Roma woman living in the district, succeeded in arguing that the principle of equal treatment applied to her. The ECJ left the actual finding of direct discrimination to the referring court, but indicated that CHEZ had indeed directly discriminated even against the complainant. As Atrey comments, this was not an instance of traditional associative discrimination, for the Court did not mind itself to draw an associative link between the complainant and Roma people.[69] Rather, the measure constituted

[66] ibid [38], emphasis added.

[67] *Coleman* (n 63), Opinion of Mr Advocate General Maduro delivered on 31 January 2008 [18].

[68] C-83/14, "*CHEZ Razpredelenie Bulgaria" AD v Komisia za zashtita ot diskriminatsia* [2015] electronic Reports of Cases.

[69] Shreya Atrey, 'Redefining Frontiers of EU Discrimination Law' (2017) PL 185, 188.

direct discrimination so long as it had been "introduced and/or maintained for reasons relating to the ethnic origin common to most of the inhabitants of the district concerned".[70] On one reading put forward by Atrey, this seems to further divorce the relationship between the characteristic and person, recognising instead a sort of "collateral discrimination"[71] claim on the complainant's part by virtue of the fact that she suffered the adverse effects of a policy that was constructed on racial stereotypes. However, a more conservative reading of the judgment is possible, limiting the width of associative discrimination. Perhaps an associative relationship existed between the complainant and Roma people because she lived in a Roma-majority district, and so she had been discriminated against for living in a district with Roma people, the subject of CHEZ's prejudice.

In domestic law, the EAT in some cases has also developed a wide view of associative discrimination. In *Showboat Entertainment Centre v Owens*,[72] the EAT held that the statutory wording "on racial grounds" in section 1 of the Race Relations Act 1976 included a case where a manager had been dismissed for refusing to obey an instruction not to allow Black people into an amusement centre. It stated that "[the] only question in each case is whether the unfavourable treatment afforded to the claimant was caused by racial considerations".[73] On this wide view, there would have been little trouble establishing associative discrimination in *Lee* because the reason why Ashers treated Lee as it did was because the requested message expressed support for *gay* marriage, factoring in sexual orientation even if only in the abstract. This was the conclusion reached by the Court of Appeal,[74] which was rejected by the Supreme Court.

The narrow view imposes a requisite connection between the recipient of the treatment and the individual(s) possessing the protected characteristic. This is the conservative reading — and Butlin argues, the "proper reading"[75] — of *Coleman*. The Court did not adopt the same exclusionary analysis as Advocate General Maduro, framing its decision more tightly around the fact that the disabled party was the dismissed woman's son. Domestically, associative discrimination began in narrow form. In *Race Relations Board v Applin*,[76] a married couple who cared for "coloured" foster children from the local authority were pressured by their neighbours to take White children only. In the Court of Appeal, Stephenson LJ concluded that "A can discriminate against B on the ground of C's colour,

[70] *CHEZ* (n 68) [3].
[71] Atrey (n 69) 188.
[72] *Showboat Entertainment Centre v Owens* [1984] 1 WLR 384.
[73] ibid 390.
[74] *Lee* (n 24) [58].
[75] Sarah Fraser Butlin, 'Cakes in the Supreme Court' (2019) 78 CLJ 280, 282.
[76] *Race Relations Board v Applin* [1973] QB 815.

race or ethnic origin".[77] In the House of Lords, Lord Simon concurred on this point, providing the example of "discriminating against a White woman on the ground that she had married a coloured man".[78] In *Applin* it seems to have been envisaged that associative discrimination would apply where there is a relationship between a third-party possessing the protected characteristic and the recipient of the treatment.

Preference for the narrow view in domestic law is still evident more recently in *Redfearn v Serco Ltd*[79] where the Court of Appeal curtailed the broad trajectory set in *Showboat*. Mr Redfearn was employed by Serco to provide transport services to customers most of whom were of Asian origin. After he was elected councillor for the British National Party, which is known for its aim of establishing a predominantly White Britain, he was dismissed by Serco on the ground that he presented a risk to the health and safety of Serco's customers and employees. The Court held that the mere fact that racial considerations had been taken into account by the alleged discriminator "[did] not mean that it is right to characterise Serco's dismissal of Mr Redfearn as being on 'racial grounds'".[80] As Forshaw and Pilgerstorfer have argued, this seems to have been more of an 'instinctive' knee-jerk reaction to Redfearn's membership of BNP, rather than a judicially reasoned decision, given the sparse legal analysis.[81] Side-by-side with *Showboat* where the claimant had refused to shut the centre to Black people, it appears the Court was teleologically interpreting discrimination law. In *Redfearn*, Mummery LJ remarked that to allow Redfearn to establish direct discrimination would be "incompatible with the purpose of the [Race Relations Act 1976] to promote equal treatment of persons irrespective of race by making it unlawful to discriminate against a person on the grounds of race".[82] In *Showboat*, Browne-Wilkinson J also sought to vindicate underlying policy, finding it "impossible to believe that Parliament intended that a person dismissed for refusing to obey an unlawful discriminatory instruction should be without a remedy".[83] Evidently, domestic courts have sensed the potential of an unreservedly wide view to distort or overstep the intended scope of direct discrimination law.

In future cases, the ambiguity in *Lee* should be resolved in favour of the narrow view. The individuals possessing the protected characteristic need not

[77] ibid 831.

[78] *Race Relations Board v Applin* [1975] AC 259, 289.

[79] *Redfearn v Serco Ltd* [2006] EWCA Civ 659, [2006] ICR 1367.

[80] ibid [46].

[81] Simon Forshaw and Marcus Pilgerstorfer, 'Taking Discrimination Personally? An Analysis of the Doctrine of Transferred Discrimination' (2008) 19 King's Law Journal 265.

[82] *Redfearn* (n 79) [46].

[83] *Showboat* (n 72) 389.

be named people known to the recipient of the treatment. So long as they are individuals belonging to a definable set, who are linked to the recipient more than simply in the alleged discriminator's abstract thought, associative discrimination should be made out. Discriminatory instruction cases such as *Showboat* would fall within direct discrimination because the employee associated with Black people by refusing to exclude them. *Redfearn* would fall outside direct discrimination because there was no set of individuals defined by race that Redfearn could be identified with for the sake of differential treatment by Serco. Conversely, Redfearn was associated with BNP, a Whites-only political group not distinguished by the White race but rather by their common political ideology that happened to be racist. There was no group of individuals defined by their race being accorded second-class citizenship or unequal dignity by the employer's acts in respect of Redfearn.

In *Lee*, there was no factual evidence that Ashers had determined its refusal on the basis of Lee's associates' characteristics. Indeed, Ashers employed gay and bisexual employees and had not discriminated against them in the past, rendering an assumption of associative discrimination rather implausible. The only way to establish associative discrimination, therefore, would be to hold that the message "Support Gay Marriage" itself yields a sufficiently close associative connection between Lee and homosexual individuals (who do not need to be determined as specific individuals, but at least as a definable set of individuals). This could not be the case because even if the message could be related to all proponents of same-sex marriage, that did not set apart a set of individuals defined by sexual orientation.[84] It is similar to *Redfearn* in that there is no set of individuals representing a characteristic that could be associated with the recipient of the treatment, except in abstract subject-matter terms. The wide view of associative discrimination would have allowed direct discrimination to be made out here simply because the abstract matter of sexual orientation was a consideration in Ashers' decision: specifically, since their belief that marriage is only between a man and a woman delineated on grounds of sexual orientation, and since it determined their treatment of Lee, direct discrimination would be established. This, however, should be rejected.

One obvious argument weighing in favour of the narrow view is legislative intention. Before the Equality Act 2010 was passed, the EAT had upheld a creative interpretation of the statute that allowed for wide associative discrimination, stating in *Zarczynska v Levy*,[85] a discriminatory instruction case, that "the strict interpretation of the relevant sections […] may well create an absurd or unjust situation which Parliament would not have intended if they had contemplated its possibility". Now, the Explanatory Notes to the Equality Act 2010 make express

[84] *Lee* (n 3) [25], [33].

[85] *Zarczynska v Levy* [1979] 1 WLR 125, 129; see Showboat (n 72) 389.

reference to associative discrimination, circumscribing it to the narrow view. It states that direct discrimination in section 13 is "broad enough to cover cases where the less favourable treatment is because of the victim's association with someone who has that characteristic".[86] The emphasis is therefore on the association between the recipient and the individual possessing the characteristic, rather than the characteristic in the abstract.

Another argument in favour of the narrow approach is normative. By unhinging protected characteristics from particular individuals, a wide view of associative discrimination would effectively prohibit certain opinions and points of view from being acted upon, which would exceed the role of the legal prohibition of direct discrimination and undermine democracy. As explained above, on a wide view, Ashers would have been regarded as unlawfully discriminating on grounds of sexual orientation by acting upon a personal belief about sexual orientation. There would have been no need to show that they perceived that Lee was a LGBTQ person, or associated with LGBTQ people. Prohibiting acts which do not fit the structure of unequal treatment of individuals because of a protected characteristic personally possessed by a relevant individual, simply because they were motivated by a belief *about* a protected characteristic, would severely affect the freedom of individuals to hold and to express beliefs about any of the protected characteristics — sex, race, gender, sexual orientation, nationality, age, and so on. In short, divorcing protected characteristics from individuals would allow discrimination law to creep outside the unequal treatment situation, into the extensive possible situations in which beliefs that have something to do with protected characteristics are acted upon in the abstract. This exceeds the mandate of the direct discrimination prohibition, which is to protect the kernel of human dignity that demands that like people be treated alike. The issue of whether the beliefs of the alleged discriminator facilitates equal human dignity or not, is a step removed from the issue of whether they treat like persons alike in their formal actions, and it is an issue meant to be determined through the political, democratic process and not the prohibition of direct discrimination.

Finally, it is worth remembering that there are other elements of the comprehensive equality legal framework operating to address these more abstract instances of discrimination. Bodies such as the Equality and Human Rights Commission are tasked to encourage the development of an equal society.[87] Analysing the *Showboat* decision, Forshaw and Pilgerstorfer point out that specific statutory protection existed against discriminatory instructions issued by employers. The statute at the time, however, only allowed such a claim to be

[86] Explanatory Notes to the Equality Act 2010, paragraph 59.

[87] Equality Act 2006, sections 1–3; Northern Ireland Act 1998, section 73.

initiated by the Commission for Racial Equality. Now, section 111(5) Equality Act 2010 would allow the manager to bring a claim with regard to the employer's instruction without having to shoehorn his case into a direct discrimination mould. Therefore, choosing the narrow view in the future would not leave a lacuna in equality protection.

E. Indirect discrimination

It has been questioned by Connolly why Lady Hale gave such short shrift to the possibility that Ashers had indirectly discriminated against Lee.[88] Observing that Brownlie J — having already found direct discrimination — did not find indirect discrimination, Lady Hale remarked that "it is not easy to see how she could have done so".[89] It is argued that indeed, *Lee* does not fit the paradigm of indirect discrimination because the particular disadvantage suffered needs to be tangible and objectively ascertainable, rather than a mere amorphous subjective impact such as offense.

To establish indirect discrimination, it must be ascertained what "particular disadvantage" has been suffered by homosexual people or supporters of gay marriage as a group.[90] In most cases, the disadvantage is tangible and objectively ascertainable. In *Bull*, for example, homosexual couples lacked the access to double rooms that heterosexual couples enjoyed. Since Ashers' policy was to refuse to ice a cake with the message "Support Gay Marriage" whatever the sexual orientation or political belief of the customer, the particular disadvantage suffered by gay people or supporters of gay marriage cannot have been their inability to obtain the cake: anyone of any characteristic would have been likewise unable to obtain it. Instead, the disadvantage can only have been a subjective impact on individuals owing to their own sexual orientation or protected belief. The most plausible expression of this subjective disadvantage for gay people is that Ashers' policy stalls the campaign for rights of equal marriage which they would personally reap. It is argued that this is too remote a link because Ashers' supply or failure to supply the cake does not directly affect whether or not gay people enjoy marriage rights. In any event, that would have appeared inconsistent with Lady Hale's later suggestion that the

[88] Michael Connolly, 'Lee v Ashers Baking and its Ramifications for Employment Law' (2019) 48 Industrial Law Journal 240, 246.

[89] *Lee* (n 3) [21].

[90] Equality Act 2010, section 19(1).

benefits of gay marriage accrue to individuals of all sexual orientations in the "wider community".[91]

It might instead be argued that gay people suffer particular offense, insult, or affront to dignity as a result of Ashers' policy, in a manner uniquely felt by them and not by heterosexual individuals. Firstly, however, this would undermine Lady Hale's central conclusion that since any customer would have been treated equally, Lee had not been treated in a degrading manner, and that therefore Ashers had not directly discriminated.[92] Secondly, this is too subjective an impact to be considered "particular" to homosexual individuals as a category. Major homosexual gay rights activists such as Tatchell have campaigned for less interference by public order legislation with "insulting" expression;[93] indeed, Tatchell supports the decision in *Lee*.[94] Further, as the Court of Appeal pointed out, "some gay people oppose gay marriage",[95] for reasons such as its perceived patriarchal legacy.[96] Founding indirect discrimination on the claim that gay people suffer special insult when service providers disagree with their right to marry would be an unwarranted assumption and would pre-empt the answer to what is actually a complex socio-political question. It would also make a significant inroad into free speech by conceding that a right not to be insulted by another person's idea horizontally competes with their freedom to express it. For these reasons, *Lee* also falls outside the limits of indirect discrimination law.

III. External Limits

In cases such as *Lee*, there arises a normative dilemma between equality law on one hand, and freedoms of religion and of expression on the other. The latter freedoms place external pressure on discrimination law, keeping it within its bounds in a liberal plural society. In this section, I examine the uniquely composite

91 *Lee* (n 3) [33].

92 ibid [35].

93 Michael White, 'Peter Tatchell's right, we don't need a law against hurt feelings' (*The Guardian*, 16 May 2012) <www.theguardian.com/politics/blog/2012/may/16/law-hurt-feelings-peter-tatchell> accessed 22 February 2021.

94 Peter Tatchell, 'I've changed my mind on the gay cake row. Here's why' (*The Guardian*, 1 February 2016) <www.theguardian.com/commentisfree/2016/feb/01/gay-cake-row-i-changed-my-mind-ashers-bakery-freedom-of-conscience-religion> accessed 22 February 2021.

95 *Lee* (n 24) [24].

96 Tom Geoghegan, 'The gay people against gay marriage' (*BBC News Washington*, 11 June 2013) <www.bbc.com/news/magazine-22758434> accessed 22 February 2021.

engagement of freedoms of religion and expression in *Lee* that compelled a different outcome from *Bull*, in which only freedom of religion had been engaged.

A preliminary structural clarification to make is that, in the Court's analysis, the Convention rights were used not as a justification *per se* of direct discrimination, as Collins has argued.[97] Instead, pursuant to the Court's duty to construe the law compatibly with Convention rights,[98] the rights "[impacted] […] the meaning and effect"[99] of the statute by determining which of the alternative interpretations of indissociability should prevail in Lee's political beliefs claim. The Court chose to follow the Convention-compatible conclusion that the criterion and characteristic were dissociable.

A. Composite engagement

Ashers' objection to the express message requested by Lee sets *Lee* apart from cases such as *Bull* and *Ladele*,[100] where there was no express message being objected to, but only an act in respect of a person with a protected characteristic that the alleged discriminator refused to carry out on grounds of conscience. *Bull*, *Ladele*, and *Lee* all involved a "balancing exercise between protection from discrimination and the rights of religious people not to be compelled to act against their conscience".[101] However, what tipped the balance in favour of Ashers in *Lee* was the crucial composite engagement of the owners' freedoms of religion and of expression.

B. Political messages concerning gay rights

Ashers' refusal to promote the message merits protection as political speech. First it must be noted that there is a distinction between protected political speech and hate speech, of which the latter falls outside the ambit of freedom of expression.

The Canadian case law in this regard, helpfully catalogued by Moon, provides examples of anti-LGBTQ religious expression where courts have sought

[97] Hugh Collins, 'A missing layer of the cake with the controversial icing' (*United Kingdom Labour Law Blog*, 4 March 2019) <https://uklabourlawblog.com/2019/03/04/a-missing-layer-of-the-cake-with-the-controversial-icing-hugh-collins/> accessed 22 February 2021.

[98] Human Rights Act 1998, section 3.

[99] *Lee* (n 3) [48].

[100] Both were decided against the religious alleged discriminators.

[101] Butlin (n 75) 283.

to draw this distinction.[102] In *Owens v Sask. (HRC)*,[103] *Lund v Boissoin*,[104] and *Whatcott v Sask. HRC*,[105] the Canadian courts protected — in the interest of free speech — newspaper ads, editorial letters, and flyers expressing the authors' opinion that homosexuality is immoral. A recurring reason for these decisions was that the expression took place amidst an "ongoing debate"[106] about the place of sexuality in Canadian policy. The expression was therefore a political contribution. Nevertheless, it must be scrutinised whether it constitutes hate speech: as Rothstein J remarked, "[speech] that has the effect of shutting down public debate cannot dodge prohibition on the basis that it promotes debate".[107] The Canadian courts have identified unlawful political *hate* speech where the speech contains "representations of detestation and vilification delegitimizing those of same-sex orientation".[108] Where political messages degrade, detest, or vilify groups in a manner that undermines their dignity, they do not merit protection.

Similarly, in the Convention context, there is a strong foundational protection of freedom of expression: Article 10 protects "not only […] 'information' or 'ideas' that are favourably received or regarded as inoffensive or as a matter of indifference, but also […] those that offend, shock or disturb the State or any sector of the population".[109] Applied to anti-LGBTQ religious expression, the Supreme Court of Sweden acquitted a pastor who had delivered a sermon that expressed critical opinions on homosexuality, on the ground that, since his speech was delivered as a church sermon, it was "not something that can be deemed to encourage or justify hatred of homosexuals".[110] The backstop for free speech was manifest hatred of homosexuals in a manner that scorned their dignity. Therefore, in *Vejdeland v Sweden*[111] the European Court of Human Rights (ECtHR) found that the conviction of a group that had distributed anti-gay leaflets in a school did not breach their Article 10 right because, the materials having manifested hatred, the interference had been "necessary in a democratic society for the protection and

[102] Richard Moon, 'Putting Faith in Hate' (CUP 2018), 121.
[103] *Owens v Saskatchewan (Human Rights Commission)* [2006] SJ N 221 (CA).
[104] *Lund v Boissoin* [2012] ABCA 300.
[105] *Whatcott v Saskatchewan Human Rights Commission* 2013 SCC 11.
[106] *Owens* (n 103) 66; *Boissoin* (n 104) 71; *Whatcott* (n 105) [200].
[107] *Whatcott* (n 105) [117].
[108] ibid [200]; see *Owens* (n 103) 62.
[109] *Handyside v United Kingdom* (1979) 1 EHRR 737, [49].
[110] *Prosecutor General v Åke Green*, Case No.B 1050-05 (29 November 2005).
[111] *Vejdeland v Sweden* App.No.1813/07 (ECtHR, 9 Feb 2012).

rights of others".[112] This was captured well by the Supreme Court in *Vejdeland*, cited by the ECtHR:

> "[The leaflets] were formulated in a way that was offensive and disparaging for homosexuals as a group and in violation of the duty under Article 10 to avoid as far as possible statements that are unwarrantably offensive to others thus constituting an assault on their rights, and without contributing to any form of public debate which could help to further mutual understanding".[113]

The objection to the message "Support Gay Marriage" in *Lee* was much milder than in these cases as it involved an omission rather than positive speech, and concerned the ancillary issue of the right to marry, not the central issue of equal dignity. It more closely approximates — and even so does not come close to the intrusiveness of — a case such as *Gündüz v Turkey*[114] where the applicant's defence of sharia law on a television debate was protected political speech despite being a religiously divisive proposal, the content of sharia even being viewed by some as socially discriminatory. To characterise Ashers' refusal as hatred would not only be inaccurate but would chill political speech concerning gay rights so severely that it would amount to censorship of dissenters.

Indeed, a look back at history will reveal the constitutional irony that would be committed if anti-LGBTQ political speech is not properly protected. Leigh argues that it would manifest the rejected Devlinist posture to moral values — but this time in favour of gay rights.[115] Lord Devlin, opposing the de-criminalization of homosexual conduct, argued that "society was entitled to enforce its shared morality over sexual conduct". Now we witness the converse development — what Leigh calls "the Devlinization of gay rights" — in instances such as the removal of a Christian student from his social work degree course because he had posted comments on social media expressing views on homosexuality and marriage.[116] If we value consistency in the enjoyment of fundamental constitutional rights regardless of a person's race, belief, sex, *et cetera*, anti-LGBTQ political statements should be protected even as they decline in popularity. This right to free speech was expressed memorably by Sedley LJ in *Redmond-Bate v DPP*: "Free speech includes not only the inoffensive but the irritating, the contentious, the eccentric, the

[112] ibid [59].
[113] ibid [15].
[114] *Gündüz v Turkey* (2005) 41 EHRR 59.
[115] Ian Leigh, 'Homophobic Speech, Equality Denial, and Religious Expression' in *Extreme Speech and Democracy* (OUP 2009) 375.
[116] *R(Ngole) v University of Sheffield* [2019] EWCA Civ 1127.

heretical, the unwelcome and the provocative provided it does not tend to provoke violence. Freedom only to speak inoffensively is not worth having".[117]

Rigorously protecting the freedom to express anti-LGBTQ views in legitimate political speech would also be favourable for the very concept of equality. Rivers has observed conceptual slippage underway in equality case law, where the notion of unequal treatment is conflated with disagreement with political ideas.[118] This would undermine the very core of the concept of equality as it was supposed to operate in a plural society, protecting the dignity of individuals amidst diversity. Instead it would be transformed into a comprehensive notion of equality that can only operate in a monolithic ideological landscape. This, Leigh has observed,[119] would run counter to our liberal visions such as Rawls' "'overlapping consensus' among people of different 'comprehensive views'" and Sachs J's statement in the South African Constitutional Court that,

> "[t]he objective of the Constitution is to allow different concepts about the nature of human existence to inhabit the same public realm, and to do so in a manner that is not mutually destructive and that at the same time enables government to function in a way that shows equal concern and respect for all".[120]

If equality slips into a concept that runs counter to these visions, it will decline in utility unless we sacrifice other essential liberal values.

C. Tacit political messages and compelled speech

Applying this protection of expression in future, courts will face the question of what activity counts as political expression so that refusal to engage in it can be regarded as objection to a certain "message" rather than to the customer's characteristic. The most obvious case is where — as in *Lee* — an express political statement such as "Support Gay Marriage" is involved. This might be complicated by the argument that a message requested by a customer cannot reasonably be attributed to the service-provider in their personal capacity.[121] It is argued, however, that it can.

Firstly, the conscience dimension of Ashers' refusal weighs heavily in favour of protecting their conduct, on the basis of their freedoms of religion

[117] *Redmond-Bate v DPP* [2000] HRLR 249, [20].
[118] Julian Rivers, 'Is Religious Freedom under Threat from British Equality Laws?' (2019) Studies in Christian Ethics 1, 11.
[119] Leigh (n 115) 396.
[120] *Minister of Home Affairs v Fourie*, Case CCT 60/04, 1 December 2005, [94].
[121] *Lee* (n 27) [67].

and expression, which are closely linked here. It has been established that certain compelled conduct such as a "caps off" instruction during Christian prayers amounts to compelled active participation in religious activity, contravening freedom of conscience.[122] Whilst courts should not be absolutely deferential in conscience claims, courts should readily respect claims such as Ashers' that participation in an expression violates their conscience, because "in its religious dimension", the conscience is "one of the most vital elements that go to make up the identity of believers and their conception of life",[123] beyond mere political convictions. Furthermore, inquiring into the plausibility of being conscience-stricken comes dangerously close to questioning the substantive reasonableness of an individual's beliefs.[124] It would perpetuate the English courts' "fairly narrow view of the salience of religion or belief in the lives of individuals" observed by Rivers, a trajectory that that may "radically […] disempower, and we might even say *outlaw* religious groups".[125]

Secondly, freedom of expression encompasses freedom not to express. Barendt has pointedly remarked that to afford individuals the freedom to speak their opinions while compelling them to speak opinions they do not hold, would be "nonsense".[126] Given the equality interest at stake for Lee, however, a preferable approach to compelled speech cases would be to weigh the effect of the compelled speech against the speaker's ability to exercise their freedom to disclaim that view in favour of their actual personal view. For, after all, the speaker retains the ability to communicate their contrary personal views in a personal capacity. Canadian and US jurisprudence have taken this approach to cases involving the payment of dues for compulsory unions and associations which then finance political activities that the union members do not personally support. In those cases, the courts have held that since the compelled payment did not prevent the complainants from personally speaking against those political activities, it did not implicate their freedom of expression.[127] Applying this approach to cases of service-providers' objections to messages, their freedom not to express should indeed be upheld. Unlike in the union cases, which only involved private payment of a fee, Ashers was asked to produce a message on a cake, which they would be known to have produced because of their branding on it. They might be regarded by some of the public as complicit in the expression of the message, and would not have many

[122] *Commodore Royal Bahamas Defence Force v Laramore* [2017] UKPC 13, [2017] 1 WLR 2752.
[123] *Kokkinakis v Greece* (1993) 17 EHRR 397 [31].
[124] *Eweida v UK* (2013) 57 EHRR 213 [81].
[125] Rivers (n 119) 5–6.
[126] Eric Barendt, *Freedom of Speech* (OUP 2007) 94.
[127] *Glickman v Wileman Bros* 521 US 457 (1997); *Lavigne v Ontario Public Service Employee Union* [1991] 2 SCR 211.

options in their personal capacity to "neutralise" or disclaim it except by positively stating their disapproval in a way that would be disproportionate and unhelpful.

Protection does not extend, however, to tacit statements represented by mere actions. In *Bull*, for example, freedom of expression was not engaged by the hoteliers' decision not to let a double room to gay couples. But the line distinguishing protected speech and mere actions must be drawn carefully. The question arose in the US Supreme Court in relation to a refusal to supply a wedding cake to a same-sex couple, in *Masterpiece Cakeshop v Colorado Civil Rights Commission*.[128] Three justices found that making a wedding cake was indeed expressive as it "celebrates a wedding, and [if it] is made for a same-sex couple it celebrates a same-sex wedding".[129] Two justices regarded a wedding cake as simply a good which is not *per se* expressive of a political idea and therefore did not merit protection.[130] This point was not pertinent in the final decision. However, it is argued here that the latter view should prevail. Especially in a politically charged climate saturated with debates ranging from investment portfolios to personal diet, nearly all conduct can be interpreted as a political message of some sort owing to their political undertones and implications. Protection of all obscurely 'political' conduct would begin to unfasten freedom of expression from its core rationales such as aiding the discovery of truth, guarding a unique channel of self-fulfilment, receiving and imparting information, and facilitating democratic discussion.[131] Nevertheless, even if the conduct complained of in *Masterpiece* did not interfere with the baker's freedom of expression, it was a claim in freedom of religion insofar as it was motivated by conscientious convictions. The claim takes on, therefore, the conscience dimension highlighted above, which has been increasingly neglected or softened in recent jurisprudence. In that regard, the Court's final decision that the Civil Rights Commission did not exhibit religious neutrality was a welcome reclaiming of conscience protection, even if the case did not finally concern freedom of expression. A balancing exercise of the actor's freedom of religion against the customer's right to non-discrimination was merited.

There remains a narrow category of 'symbolic speech' — or 'expressive conduct'— that constitutes protected expression even though it does not involve an express message. Examples include saluting a national flag,[132] taking one's cap off,[133] or conducting a silent sit-in.[134] These are forms of conduct, supported by deep cultural

[128] *Masterpiece Cakeshop v Colorado Civil Rights Commission* 138 SCt1719 (2018).
[129] ibid 1738.
[130] ibid 1751.
[131] Barendt (n 127) 2.
[132] *West Virginia State Board of Education v Barnette* 319 US 624 (1943).
[133] *Commodore Royal Bahamas Defence Force* (n 122).
[134] *Brown v Louisiana* 86 S.Ct.719 (1966).

history, that can plausibly be understood as "communicative"[135] of a message: a silent gesture delivering assent or dissent in the same way a spoken message would. In contrast, the provision of a cake, without the element of a message, or hotel room, is primarily the provision of a service to other persons.

It is worth noting, to close, that the judgment in *Lee* was handed down in a plural and diverse society, balancing the values that compete uniquely in our current social context. The most democratic way to foster equality and mutual respect in our diversity is not by prohibiting expression of views that are thought to be illiberal, but rather, as Geddis argues, by a more "transformative" strategy whereby "the public [learns] to tolerate […] offence in the name of a vibrant, robust and open realm of public discourse".[136] The outcome in *Lee* might have been different in a less diverse society where, as Knights hypothesises, the allegedly discriminatory "service providers […] effectively have monopolies" or "minority views […] are widely opposed".[137] As it happens, Lee managed to obtain the cake elsewhere. But if, in that hypothetical society, every bakery had refused Lee's order, then Lee would have not been able to obtain his desired cake at all. If such a homogeneous society was the backdrop of Ashers' conduct, then the service-providers' freedoms of religion and expression should have been more readily limited in favour of equality and freedom of expression for Lee. This different balance would have been achieved with flexible proportionality analysis.[138]

IV. Conclusion

In this article I have sought to mark the internal and external limits of discrimination law confronted by the Supreme Court in *Lee*. I have expanded on the brief treatment given in the judgment to the tools of discrimination law and Convention rights, which are far more complex than meets the eye. The internal tools must be carefully applied to preserve the core aim of discrimination prohibitions, that is, the prohibition of differential treatment of individuals with protected characteristics. Viewed on an analytical level, courts must carefully guard the requisite link between treatment afforded and the protected characteristics. The guaranteed freedoms of alleged discriminators place external pressure on discrimination law, which generally interferes with individual autonomy and

[135] *Clark v Community for Creative Non-Violence* 104 SCt3065 (1984).

[136] Andrew Geddis, 'Free Speech Martyrs or Unreasonable Threats to Social Peace?—"Insulting" Expression and Section 5 of the Public Order Act 1986' (2004) PL 853.

[137] Samantha Knights, 'Case Comment: Lee v Ashers Baking Company Ltd & Ors.' (*United Kingdom Supreme Court Blog*, 12 November 2018), <http://ukscblog.com/case-comment-lee-v-ashers-baking-company-ltd-ors-2018-uksc-49-2/> accessed 22 February 2021.

[138] See *Bull* (n 2) [45].

specifically interferes with freedoms of religion and expression. To guard the diversity and pluralism that we value as a society, these freedoms must be protected even in respect of views with which we disagree, provided they are not violent or hateful. Whilst tracing these limits and identifying the values vindicated by them, I have also proposed trajectories for their future application. In sum, it is argued that the decision was a welcome bridling of discrimination law, an area in which expansions can be tempting owing to the nobility of the aim of equality, but which must be limited for the sake of other liberal values.

Cambridge Law Review (2021) Vol VI, Issue i, 236–262

What Happens in the Jury Room Stays in the Jury Room: *R v Mirza*, the Criminal Justice and Courts Act, and the Problem of Racial Bias

Nicholas Goldrosen*

Abstract

This article argues that courts' refusal to consider juror testimony about deliberations and the laws restricting jurors from speaking about deliberations prevent defendants from seeking adequate redress for juror racial bias. The article first presents a brief history of the common law and statutory foundations of jury secrecy under English law. I then argue that juror racial bias uniquely threatens the right to an impartial tribunal and that other safeguards are not necessarily adequate to ameliorate or prevent bias during deliberations. English courts have historically upheld jury secrecy by holding that the interests of finality and candour outweigh the injury done to a defendant by juror racial bias, as exemplified in *R v Mirza*. While the Criminal Justice and Courts Act 2015 does make some changes to jury secrecy law — mainly by allowing jurors to report some forms of misconduct that occur during deliberations — this article argues that the Act inadequately protects defendants. The Act's reporting provisions are overly complex, largely non-adversarial, and too focused on enabling the prosecution of jurors who commit misconduct. I argue that a reform of this Act to more explicitly focus on

* MPhil Candidate, Institute of Criminology, University of Cambridge. BA (Williams College). I am grateful to the editors of the Cambridge Law Review for their helpful comments. ncg36@cam.ac.uk.

protecting defendants from juror misconduct — and in particular, juror racial bias — is necessary to better secure defendants' fair trial rights.

Keywords: law of juries, criminal procedure, racial bias and the law, jury secrecy, fair trial rights

I. Introduction

The pithy tourism slogan for the city of Las Vegas — "what happens in Vegas, stays in Vegas"[1] — would be an apt slogan for the English jury room, too. The criminal law, of course, relies on various forms of compartmentalisation and regulated disclosure of information. For example, judges prevent juries from considering inadmissible evidence, so that the jury might see only what is legal and relevant. With regards to juries, the English law restricts everyone, including the courts themselves, from examining what occurs when the jurors retire and deliberate.[2] Jury secrecy in the English legal system is maintained by two legal instruments — a prohibition on inquiry into the jury room and a prohibition on speaking out from the jury room. The former, Mansfield's rule, prohibits courts from considering juror testimony to undermine or overturn a conviction. The courts cannot, in a legal sense, 'see' what occurs in the jury room. As to the latter, jurors face criminal penalties they face if they disclose their deliberations. The jury room becomes a space set apart, legally speaking.

This secrecy becomes problematic when something goes awry in the jury room. After all, one important thing that happens in the jury room does not stay inside of it — the verdict. A jury's decision-making, though secretive and based on a limited universe of information, has real consequences for the defendant. If a juror commits misconduct, the available remedies are limited by jurors' inability to report the issue and courts' inability to grant relief based on juror testimony. Given the historic and present racial injustices in the criminal legal system, one of the most concerning scenarios is when a juror makes racially prejudiced remarks during deliberations. This type of bias is anathema to the impartial tribunal to which all defendants are entitled, but jury secrecy obstructs the court from giving

[1] Samantha Shankman, 'A Brief History of "What Happens in Vegas Stays in Vegas' *The Week* (New York, 1 October 2013) https://theweek.com/articles/459434/brief-history-what-happens-vegas-stays-vegas accessed 3 May 2021.

[2] With apologies to the Welsh, I use 'English legal system' and variants thereof throughout this Article to refer to the unified criminal legal system of England and Wales. While the Welsh government has some devolved powers with regard to criminal justice, it does not have such powers over any matters at issue in this Article, such as juries or criminal procedure.

the defendant any relief. The law on jury secrecy is constructed so as to wilfully obscure racial bias from the view of the courts.[3]

This article explores the conflict of jury secrecy and racial bias, particularly through the lens of the Criminal Justice and Courts Act 2015 and *R v. Mirza*.[4] The latter was a 2004 case before the House of Lords, in which the Law Lords refused to admit evidence of a juror's racial bias to quash the defendant's conviction. The former went part way toward removing the gag from jurors; the Act created some exceptions, in the event of juror misconduct, to the blanket bar on jurors' disclosure of deliberations. The Act, though, is insufficient. Intended mainly as an instrument to prosecute jurors for misconduct, the Act does not include racial bias amongst its exceptions. Furthermore, its process for reporting misconduct is unwieldy, non-adversarial, and would have a chilling effect on jurors reporting misconduct.

Section II of this article presents a brief history of juror secrecy and Mansfield's rule. These features are not necessarily as long-standing a part of the English legal system as they are often portrayed, but they nonetheless have found broad purchase in the courts. Section III analyses the conflict of racial bias and jury secrecy in *Mirza.* Racial bias is a paramount threat to the jury system — both to the rights of the individual defendant and the legitimacy of the system as a whole. Section IV catalogues the accomplishments and shortcomings of the 2015 Act. In particular, the Act fails to create a workable structure for handling the types of jury bias present in *Mirza*. Section V proposes appropriate reforms to the 2015 Act, balancing the need for finality and confidentiality in jury verdicts with the guarantee of an impartial, unbiased jury. Few, if any, cases have dealt with the Act's jury secrecy provisions; this area of law is ripe for legislative intervention to strengthen it before another wrenching test case of racial bias occurs. If the English justice system truly intends to treat each defendant without fear or favour, it must open the door to the jury room at least a crack; only then can defendants gain adequate relief when racial bias taints their convictions.

II. A Brief History of Jury Secrecy

A. Statutory jury secrecy

English jury secrecy is a product of both common law and statute, which operate in tandem. As to the latter, a number of statutes serve to prevent jurors from disclosing information from deliberations. The former is applied not to the jurors but to the courts; they are barred by common law from considering juror

[3] While I refer throughout to 'racial bias,' the Equality Act of 2010 includes discrimination based upon ethnicity, nationality, national origin, and colour as racial discrimination.

[4] *R v Mirza* [2004] UKHL 2.

testimony to quash a conviction. Thus, the courts must wilfully refuse to consider some forms of misconduct that might occur in the jury room. I will first detail the history of statutory jury secrecy in modern English law; the next section will consider the common law basis for excluding juror testimony.

The proceedings of English juries are secret; no information from their deliberations (save the verdict, of course) may be disclosed by the jurors nor solicited from them. While courts often cite this principle as a long-standing part of the common law, its actual origins are unclear.[5] The recent modern history of the principle in common law and statute is well-documented. As to the common law basis, in the 1962 case *R v Thompson,* Lord Parker held for the Court of Appeal that jury deliberations were to be secret and not enquired into by the court, nor anyone else.[6] Some 19 years later, Parliament enacted the Contempt of Court Act 1981, providing a statutory basis for the secrecy of jury deliberations. The 1981 Act barred divulging jury deliberations, making it a criminal offence to "obtain, disclose or solicit any particulars of statements made, opinions expressed, arguments advanced or votes cast by members of a jury in the course of their deliberations in any legal proceedings".[7] The Act provided for imprisonment of up to two years as a penalty.[8]

The practical limits of this secrecy have been tested most acutely in instances of juror misconduct. This gag rule also hinders the government's ability to prosecute and combat other forms of juror misconduct, as jurors cannot reveal what occurred in deliberations. In some cases, where juror misconduct has occurred outside of the deliberation room, courts have held that disclosing and considering this information is not barred.[9] English courts have repeatedly held that misconduct occurring inside the jury room could not be divulged outside it, though. The Act did, in theory, make a small exemption: Testimony that encompassed jury deliberations could be used in giving evidence for an "offence committed in relation to the jury".[10] In practice, this exception was usually construed so narrowly as to be useless; the exception allowed for jurors to give this information in court, but

5 Pamela Ferguson, 'The Criminal Jury in England and Scotland: The Confidentiality Principle and the Investigation of Misconduct,' (2006) 10 International Journal of Evidence & Proof 180, 181. This article provides an excellent treatment of the state of jury secrecy law in 2006 along with a comparative treatment of English and Scots law.

6 *R v Thompson* [1962] 4 Cr App R 72.

7 Contempt of Court Act 1981, section 8

8 ibid, Section 14.

9 *R v Young* [1995] QB 324. In this case, the court held that evidence of jurors engaging in a séance to aid in their deliberations was admissible, because it occurred in a hotel, not in the deliberation room. For a good exposition of this case, and of other jury secrecy rulings, see generally Lord Robert Reed, 'The confidentiality of jury deliberations' (2003) 37 The Law Teacher 1.

10 Contempt of Court Act 1981 (n 7) [8].

there was no provision for them to ever notify anyone about the jury misconduct in the first place. For example, in *Attorney General v Scotcher*, the House of Lords upheld the conviction of a juror who told the mother of the defendant information that he believed revealed juror misconduct.[11] If jurors could not notify anyone of misconduct after a verdict without facing prosecution, it made little difference that the 1981 Act allowed for them to give evidence in criminal proceedings regarding that misconduct. The seal of the jury room made sure that misconduct would remain unknown.

B. Mansfield's rule

Of course, the ability of jurors to speak about misconduct would only be a partial step towards remedying that misconduct; it is also necessary that the court be willing to consider that testimony and provide relief to the defendant. Yet the English legal system has been deeply resistant to allowing jurors to impeach their own verdicts via affidavit or testimony. In the 2010 case *R v Thompson*, for example, the Court of Appeal held that it could not admit any evidence of juror misconduct involving internet research during deliberations, even if jurors would not face prosecution for disclosing that information.[12] The common law rule against juror testimony prohibits courts from 'seeing', in a legally meaningful way, juror misconduct that might threaten a conviction.

This rule's current form follows from the judgment of Lord Mansfield in two late 18th-century cases, *R v Almon* and *Vaise v Delaval*, in which he barred juror testimony that the verdict had been reached by a game of chance.[13] This rule has hence sometimes been termed 'Mansfield's Rule,' particularly in the United States of America, where it was adopted in the Federal Rules of Evidence.[14] Yet the actual historical roots of this practice prior to 1785 are murky. Several cases decided by English courts in the years prior to *Vaise* held that juror testimony about various forms of misconduct during deliberations could be admitted to impeach a conviction.[15] For example, in *Metcalfe v Deane*, the court admitted evidence concerning the jurors undertaking investigation and questioning on their own.[16] In *Prior v Powers*, the court similarly heard evidence that the jurors arrived at a decision

[11] *Attorney General v Scotcher* [2005] UKHL 36.

[12] *R v Thompson* [2010] EWCA Crim 1623.

[13] *R v Almon* [1770] 98 ER 411 (KB) and *Vaise v Delaval* [1785] 99 ER 944 (KB).

[14] US Federal Rule of Evidence 606(b).

[15] Andrew Hull, 'Unearthing Mansfield's Rule: Analyzing the Appropriateness of Federal Rule of Evidence 606(b) in Light of the Common Law Tradition' (2014) 38 Southern Illinois University Law Review 403, 411–412.

[16] *Metcalfe v Deane* [1590] 78 ER 445 (QB) 445 cited at Hull (n 15) 411.

via coin toss.[17] Hence, the historical precedent for the common law's exclusion of juror testimony is not so certain.

Nonetheless, the inadmissibility of juror testimony has remained constant in modern English law. The usual justification for this testimony's inadmissibility is two-fold. First, the finality of a jury's verdict is valuable. Admission of testimony from a small number of jurors afterwards would consistently throw criminal convictions into doubt; such a problem could be especially bad following convictions by majority verdict, where a disgruntled juror in the minority could assail the verdict. The courts, the government, and the victim all benefit from convictions being final. *R v Qureshi* affirmed the particularly sacrosanct nature of jury deliberations after a verdict has been reached, given the emphasis on finality.[18] Secondly, the admissibility of juror testimony would incentivise defence counsel and other parties to place pressure upon jurors to reveal their deliberations in hopes of finding a ground for appeal. This pressure would in turn undermine the candour of jurors during deliberations. In *Ellis v Deheer*, Lord Justice Atkin cited these two reasons for refusing to admit juror affidavits of misconduct, writing,

> "to my mind it is a principle which it is of the highest importance in the interest of justice to maintain, and an infringement of the rule appears to me a very serious interference with the administration of justice".[19]

The European Court of Human Rights further held in *Gregory v United Kingdom* that the inadmissibility of juror testimony was a "crucial and legitimate feature of English trial law",[20] and allowable under European human rights laws. Finally, restrictions on post-conviction investigation in other jurisdictions, such as the U.S., also cite these same rationales.[21]

Court judgments usually refer to jury secrecy as an established part of the common law, rather than statute. In *R v Andrew Brown*, which the Lords cited in *Mirza*, the Supreme Court of New South Wales held that juror testimony would be always inadmissible.[22] Chief Justice Darley wrote, "I have come to the conclusion that the authorities are all one way, and that the Court cannot look at the affidavits of jurymen for any purpose, whether it be for the purpose of granting a new

[17] *Prior v Powers* [1734] 94 ER 993 (KB) 993 cited at Hull (n 15) 412.
[18] *R v Qureshi* [2002] 1 WLR 518.
[19] *Ellis v Deheer* [1922] 2 KB 121.
[20] *Gregory v United Kingdom* [1997] 25 EHRR 577 [44].
[21] Kathryn E Miller, 'The Attorneys Are Bound and the Witnesses Are Gagged: State Limits on Post-Conviction Investigation in Criminal Cases' (2018) 106 California Law Review 160.
[22] At the time, New South Wales was under British law and jurisdiction, with Australia gaining independence in 1942.

trial, or for the purpose of establishing the misconduct of a juryman".[23] Even after jury secrecy laws were passed in the late-20th century, the Court of Appeal found that juror testimony is inadmissible under the common law for the purposes of establishing juror misconduct.[24] Lord Justice Kennedy wrote for the court in *Miah and Akbar*, "the barrier to the reception of material is not to be found in the 1981 Act. It is to be found in a long line of authorities".[25] Thus, the common law prohibition of admission of juror evidence to impeach convictions appears resistant to modern legal challenges. While its well-established nature is often cited by courts to support the rule, some pre-modern cases call into question whether it has always existed in its present form.

The traditional rationale for jury secrecy has been that the harm of breaching that of secrecy outweighed the injury suffered due to undiscovered juror misconduct. Some later statutes, nevertheless, do begin to recognise the harm that jury misconduct can do. For example, the 1981 Act, by allowing prosecutions for juror misconduct, allows the state to seek remedy when its interests are harmed by juror misconduct. When an offence is committed against the jury, however, both the state and defendant are harmed; the defendant suffers from the loss of an impartial tribunal deciding the case solely based on the evidence. Yet the 1981 Act solely provides relief to the state. Mansfield's Rule denies relief to the defendant by disallowing the courts from considering juror testimony to overturn the verdict. This dearth of remedies for the defendant following juror misconduct is the crux of the issue surrounding juror racial bias.

III. The Problem of Racial Bias and Jury Secrecy

A. Bias runs up against secrecy: *Mirza*

Racial bias on the part of jurors poses a particularly thorny challenge to both juror secrecy and the admissibility of juror testimony. A defendant tried by jury has the right not only to a jury that is correct in its composition and deliberates on the evidence to reach a verdict, but also to an impartial jury.[26] The impartial jury must be free from personal prejudices or biases.[27] Yet the absolute secrecy of the jury room shelters whatever biases might arise during deliberations from the scrutiny of the court. Indeed, *Qureshi*, the case that held juror testimony to

[23] [1907] 7 NSW State Reports 299.
[24] *R v Miah and Akbar* [1996] EWCA Crim 1653.
[25] ibid.
[26] *Ferguson* (n 5) 188.
[27] *Pullar v UK* [1996] ECtHR 23.

be always inadmissible, concerned racism in the jury room.[28] This conflict again shone through in the case of *R v Mirza*, which was considered before the House of Lords.[29] In this case, a Pakistani-British man appealed his conviction for several sexual offences against children on the grounds that a juror's racial bias against him tarnished the conviction. Specifically, he contended that the jury drew adverse inferences against him for his use of an interpreter and used racially biased reasoning and language in doing so. The House of Lords held that the rules of jury secrecy must prevail and that this evidence would be inadmissible to quash his conviction. Both the majority and dissent frame the question of *Mirza* similarly: *Does the threat of racial bias outweigh the public policy interests served by jury secrecy, given other methods of reducing jury bias?*

The majority's balancing act came down on the side of jury secrecy, finding that it outweighed the defendant's interests in admitting the juror testimony. The Lords in the majority began by noting the great consideration usually given to the historic justifications for jury secrecy: finality and candour of deliberations. Referring to these two factors, Lord Slynn held, "if there can be a review of what happens between jurors, whether in the jury box or in the jury room, the advantages relied on as justifying the rule will disappear or fundamentally be diminished".[30] Lord Slynn added another connected justification for jury secrecy, too — public confidence. He expressed concern that allowing such testimony would give rise to false or spurious allegations, undermining the jury system.[31]

On the other side of this balancing test, the majority argued that safeguards such as random selection would help minimise bias and made the Lords' ruling compatible with Article 6 of the European Convention on Human Rights.[32] Thus, they argued that juror bias could be adequately combatted without enquiring into juror deliberations. Lord Hope additionally argued that, in other jurisdictions which had applied this balancing test, the interest in juror secrecy outweighed even egregious juror misconduct. He cited an American case where the U.S. Supreme Court refused to admit evidence of jurors using alcohol, marijuana, and cocaine during the trial and deliberations.[33] The majority in *Mirza* concluded that the public policy implications of Mansfield's rule such as finality, candour of deliberations, and public confidence outweighed the threat of racial bias, given the

[28] *Qureshi* (n 18).
[29] *Mirza* (n 4).
[30] ibid [52].
[31] ibid [53].
[32] Gillian Daly, 'Jury Secrecy: *R v Mirza*; *R v Connor and Rollock*' (2004) 8(3) The International Journal of Evidence & Proof 186, 188.
[33] *Tanner v United States* 483 US 107 (1987) cited at *Mirza* (n 4) [98].

other methods judges could use to prevent such bias. This balancing act kept the jury room as a chamber of secrets.

Lord Steyn's dissent in *Mirza* ostensibly rejected this balancing act framework, voting instead to grant Mirza's appeal and quash the conviction. Lord Steyn wrote of balancing jury secrecy and the right to an impartial jury,

> "one is not dealing with a cost/benefit analysis: a miscarriage of justice bears on real individuals, their families, and communities. If the law requires individual cases to be subordinated to systemic considerations affecting the jury system, one may question whether the law has not lost its moral underpinning".[34]

Lord Steyn's assertion, however, was not wholly correct; he was, indeed, doing something of a cost/benefit analysis, but he viewed the costs of racial bias as simply too high to ever justify the benefits of secrecy. Stressing the crucial importance of an impartial tribunal as a matter of "elementary law", he argued that ignoring a blatant instance of racial bias amongst jurors, and refusing to even investigate further, would undermine this right and public confidence in the jury system.[35] The 'elementar[iness]' of the principle of impartiality, in Lord Steyn's dissent, would overpower state's interest in finality. Ultimately, both the majority and Steyn identify the same key principles in assessing the case — the right to an impartial jury, the policy implications of secrecy, and alternatives to correcting bias — but resolve the balancing act differently. Indeed, these same factors reoccur in this article's analysis of the Criminal Justice and Courts Act 2015 in sections IV and V.

B. Why does racial bias threaten juries?

(i) Bias Undermines the Impartial Tribunal and the System's Legitimacy

Racial bias denies the defendant a fair trial, a fundamental right protected under Article 6 of the European Convention on Human Rights and thus incorporated into UK law.[36] In the case *Sander v United Kingdom*, the European Court of Human Rights considered whether a judge's admonition to the jury in the summing-up was enough to ensure a fair trial after a juror reported other jurors making racist jokes. [37] The court concluded that it was not; trial by such a

[34] *Mirza* (n 4) [5].
[35] ibid [5].
[36] Human Rights Act 1998.
[37] *Sander v United Kingdom* [2000] ECtHR 19.

partial jury violated the Convention.[38] In its judgment, the court noted, "[i]t is of fundamental importance in a democratic society that the courts inspire confidence in the public and above all, as far as criminal proceedings are concerned, in the accused".[39] Insofar as the jury was biased, this trust was grievously undermined; not only the defendant, but the legal system as a whole, was harmed.

Racial bias should be of special concern to criminal justice policymakers. Black people and other racial and ethnic minorities are disproportionately represented amongst defendants and prisoners in the English criminal justice system.[40] The racial disproportionality in the English prison system has, at times, outpaced even that of the United States.[41] Many of the moral panics over crime in British society are intrinsically tied up with anti-Black racism.[42] Racial bias is a unique problem for juries because it is a uniquely harmful problem for the entire criminal legal system.

The role of the jury, however, vis-à-vis racial bias is complex. On the one hand, juries — even all-white juries — tend to convict defendants of all races at relatively similar rates.[43] Conversely, a body of evidence in social psychology indicates that negative stereotypes associating racial and ethnic minorities with criminality are common throughout the population.[44] One study using mock civil juries in the U.S. showed that jurors perceived Black defendants as more culpable, especially those with stronger accents.[45] The problem, though, is that no one has any sense of the scope of this problem because jury deliberations are so secretive. Indeed, the Contempt of Court Act prohibits any research into actual jury decision-making; the research on racial bias conducted thus far has only used mock juries.[46] At the same time, the perceived legitimacy of the jury system, especially in the eyes of citizens from marginalised racial and ethnic groups, does not necessarily concord with empirical findings on jury verdicts. Trust in juries is high generally

38 ibid.

39 ibid [22].

40 David Lammy, 'The Lammy Review: An independent review into the treatment of, and outcomes for Black, Asian and Minority Ethnic individuals in the Criminal Justice System' (8 September 2017) < https://www.gov.uk/government/publications/lammy-review-final-report>.

41 Michael Tonry, 'Racial Disproportion in US Prisons' (1994) 34 British J of Crim 97.

42 Coretta Phillips and Ben Bowling, 'Ethnicities, Racism, Crime, and Criminal Justice' in Alison Liebling, Shadd Maruna, and Lesley McAra (eds.), *The Oxford Handbook of Criminology* (Oxford University Press 2017) and Paul Gilroy, 'The Myth of Black Criminality' (1982) 19 Socialist Register 46.

43 Cheryl Thomas, 'Are Juries Fair?' (2010) 1/10 Ministry of Justice Research Series, 16.

44 Gillian Daly and Rosemary Pattenden, 'Racial Bias and the English Criminal Trial Jury' (2005) 64(3) Cambridge LJ 678, 681.

45 Jason A Cantone et al., 'Sounding Guilty: How Accent Bias Affects Juror Judgments of Culpability' (2019) 17 Journal of Ethnicity in Criminal Justice 228, 245.

46 Daly and Pattenden (n 44) 679.

across the population, but it is typically higher amongst White people.[47] That jury research has shown that most juries are fair does not mean that the legitimacy ascribed to them reflects that.

Additional safeguards would vitally improve trust in juries, particularly amongst racial groups that disproportionately seek jury trials. Black and Asian defendants are more likely than white defendants to plead not guilty and have their case tried by a jury in Crown Court.[48] Thus, maintaining confidence in the jury system amongst Black and Asian communities is paramount. Even one instance of racial bias deprives the defendant of an impartial jury. Though some reports of racial bias might be false or frivolous, they must all be dealt with because any given allegation might be true and could do incredible harm to that defendant; the possibility of this costly and time-consuming process is the only choice that ensures the impartiality of the jury.[49] Additionally, if such an accusation were made about a judge or magistrate, it would be thoroughly investigated.[50] To suspend such investigation for a jury seems inconsistent, especially given that judges and juries are governed by the same legal standard for impartiality.[51] The benefits of the improved legitimacy of the system, and remedying even a minimal number of

[47] Valerie P. Hans, 'Jury Systems Around the World' (2008) 305 Cornell Law Faculty Publications Paper 276, 282.

[48] Cheryl Thomas, 'Ethnicity and the Fairness of Jury Trials in England and Wales 2006-2014' (2017) 11 Crim LR 860, 867. This fact is somewhat difficult to square with the higher trust in juries amongst White people; what this statistic might indicate is that, while people from racial and ethnic minority backgrounds trust juries less, they trust other criminal justice players such as judges even less than that—and hence opt for trial by jury more often.

[49] John Spencer, 'Did the Jury Misbehave? Don't Ask, Because We Do Not Want to Know' (2002) 61(2) Cambridge LJ 291, 293.

[50] Daly and Pattenden (n 44) 696.

[51] *R v Brown* [2001] EWCA Crim 2828.

wrongful convictions, are justification enough for increasing scrutiny on racial bias amongst juries.[52]

(ii) Pre-verdict Safeguards Against Bias Are Inadequate

Of course, the jury system has numerous safeguards besides post-verdict investigation that are intended to guard against a racially biased conviction. Most of these protections do not offer adequate assurance of an impartial jury, though. The chief four safeguards include the jury oath, random jury selection, unanimous decision-making, and the ability to report misconduct to the judge prior to a verdict.[53] Taking these in turn, a momentary oath is unlikely to change long-held prejudices. In *Sander*, the Strasbourg Court wrote, referencing later warnings from judges, "generally speaking, an admonition or direction by a judge, however clear, detailed and forceful, would not change racist views overnight".[54] Recent psychological work shows that even longer trainings on implicit bias likely lose any effect after a few hours.[55] Hence, oaths or warnings are not likely to actually reduce juror bias.

Random jury selection is theoretically a good way to ensure a diverse set of jurors, including with regard to race. But random selection might just as easily include a racially biased juror. As Darbyshire argues, "random selection may throw up juries which are all male, all Conservative, all white".[56] Without the ability to question jurors on their views, defendants have no safeguard to ensure a racist

[52] These concerns about racial bias and juries have focused on wrongful convictions; racial bias might flow the other way, too, leading to perverse acquittals of groups for which a juror has a favourable prejudice. One rather practical reason for the focus on wrongful convictions is that the possibilities for appealing a wrongful acquittal are much narrower than for appealing a wrongful conviction, given the limited circumstances for re-prosecution under the Criminal Justice Act 2003. Hence, courts have not had to contend with a prosecutor appealing a wrongful acquittal on the ground of positive racial bias. The second, more normative, argument is that a wrongful acquittal is less severe a miscarriage of justice than is a wrongful conviction. This argument stems from Blackstone's ratio — better 10 guilty people be let free than one innocent person be convicted. The legal system has a broad and long-standing commitment to weighing the risk of wrongful conviction more heavily than that of wrongful acquittal.

[53] Daly and Pattenden (n 44) 685.

[54] *Sander* (n 37) [30].

[55] Calvin Lai et al., 'Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time' (2016) 148(8) J Exp Psychol Gen 1001.

[56] Penny Darbyshire, 'The Lamp That Shows That Freedom Lives – Is It Worth The Candle?' [1991] Crim LR 740, 745.

juror is not seated.[57] Furthermore, defendants have no right to a racially diverse jury, as the Court of Appeal held in *R v Ford*.[58]

The requirement that the jury reaches a unanimous verdict might also be a helpful safeguard against the prejudices of a majority. The introduction of majority verdicts on English juries undermines this safeguard.[59] One or two individuals who object to a biased verdict could be overruled by the rest of the jury. A majority verdict is of particular concern in the trial of a racial or ethnic minority defendant, because it could exclude a small number of racial and ethnic minority members on the jury, essentially allowing an all-white majority to decide the verdict.[60] Of course, the counterargument is that majority verdicts allow the jury to exclude the voice of a small number of biased jurors and that the majority verdict is itself a safeguard.[61] In any case, it would be foolish to rely on either unanimous or majority verdicts to protect from racial bias in every case — both have their flaws. The manner of voting cannot scrub bias from that decision-making process.

Finally, no jury secrecy law precludes jurors from drawing the judge's attention to misconduct or bias before a verdict is rendered. Judges might then warn or discharge the jury. Certainly, it is preferable that misconduct or bias be reported immediately, and the jury discharged, saving the defendant from a wrongful conviction and appeal. The costliness and time-intensiveness of a new trial often mean that a warning is given instead, though. As mentioned earlier in the *Sander* ruling, a simple warning is hardly effective at overcoming a biased jury.[62] Additionally, jurors often fail to report misconduct during the case or deliberations. They might fully realise its implications only after the verdict is rendered or feel pressure not to report the misconduct of a fellow juror in court. *Sander* is a good example here, too: The juror who initially reported the misconduct was clearly identified to the rest of the jury and was then pressured by them to sign a letter recanting the allegation of bias.[63] While reporting misconduct before the verdict is preferable for all parties, in reality, jurors often report afterwards. Even when they do report bias during the trial or deliberations, the common remedy of a warning

57 Daly and Pattenden (n 44) 685.

58 *R v Ford* [1989] 89 Cr App R 278.

59 Criminal Justice Act 1967, section 13.

60 Indeed, the United States Supreme Court ruled in *Ramos v Louisiana* 590 US ___ (2020) that non-unanimous jury verdicts violate a defendant's constitutional right to a jury trial. Majority verdicts had been used for a long time as a way to functionally exclude Black jurors, particularly in Southern states, by allowing 10 White jurors to decide the case (2).

61 Daly and Pattenden (n 44) 687.

62 *Sander* (n 37) [30].

63 ibid [29].

is insufficient. None of these other safeguards are adequate to address juror bias that is revealed after the verdict.

The balancing test employed in *Mirza* weighed jury secrecy against the safety of the defendant's conviction. Contrary to the Lords' majority in that case, however, racial bias so heavily threatens the integrity of both an individual conviction and the legal system as a whole that it outweighs the public policy interests in secrecy. Furthermore, the other safeguards in which the *Mirza* majority places faith are ineffective remedies for juror bias. *Mirza*'s balancing test was incorrect in that it failed to give enough weight to the threat of racial bias and gave too much weight to these alternatives to admitting juror testimony. The next section will argue that the Criminal Justice and Courts Act 2015, though an improvement, still fails to properly assess and combat the threat of racial bias.

IV. The Criminal Justice and Courts Act 2015 Falls Short

A. The 2015 act: a solution to juror misconduct?

The Criminal Justice and Courts Act 2015 allowed jurors to speak about deliberations in a few instances and created statutory exceptions to juror secrecy with regard to juror misconduct. This Act ostensibly filled part of the great legal void left by *Mirza*, *Scotcher*, and other cases. Even prior to *Mirza*, legal commentators had identified the lack of a procedure for investigating juror misconduct as an issue. Lord Justice Auld, in his 2001 review of the criminal courts, advocated for amending the Contempt of Court Act 1981 to allow for judicial investigation into "alleged impropriety by a jury, whether in the course of its deliberations or otherwise".[64] The Law Commission, in a 2013 report, also called for Parliament to create a limited exemption from prosecution for jurors to discuss their deliberations in reporting misconduct or a miscarriage of justice.[65] The Criminal Justice and Courts Act partially met these demands by creating exactly that limited manner of reporting.

The 2015 Act addressed a broad range of topics, touching on both civil and criminal proceedings. Its provisions on juries are focused on jury secrecy, though the Act did also raise the maximum eligibility age for jury service to 75.[66] The Act's jury secrecy provisions can be grouped into four categories: new regulations about electronic devices, new jury misconduct offences, exceptions to the bar on disclosing deliberations, and disqualification of those committing jury offences.

[64] Lord Justice Robin Auld, *Review of the Criminal Courts of England and Wales: Report* (2001) 173.

[65] Law Commission, Contempt of Court (1): Juror Misconduct and Internet Publications (Law Com No 340, 2013) 95.

[66] Criminal Justice and Courts Act 2015, Section 68.

The first and final of those categories enacted fairly straightforward changes: Judges can now order jurors, in certain instances, to surrender electronic devices and enlist court officers in enforcing that prohibition.[67] Additionally, jurors who are convicted of misconduct will be barred from jury service for 10 years.[68]

The main portion of the 2015 Act concerning juries is devoted to amending the Juries Act 1974 to create new offences for juror misconduct and exceptions to those offences. Sections 71 and 72 of the Act make it a crime under the Juries Act — rather than simply a form of contempt under the common law — for jurors to conduct their own research or share it with other jurors.[69] Section 73, somewhat tautologically, prohibits jurors from engaging in "prohibited conduct". Offering a definition, the law clarifies that, "'prohibited conduct' means conduct from which it may reasonably be concluded that the person intends to try the issue otherwise than on the basis of the evidence presented in the proceedings on the issue".[70] Finally, Section 74 makes it an offence under the Juries Act to disclose the content of jury deliberations.[71] Each of these crimes is punishable by a fine, up to two years' imprisonment, or both. Section 74 also creates a list of exceptions under which jurors may disclose deliberations.

In these exceptions, the Act lays out both the acceptable reasons for disclosure and the people to whom jurors may disclose information. The two reasons for which a juror may disclose deliberations under Section 74 are if

> "an offence or contempt of court has been, or may have been, committed by or in relation to a juror in connection with those 'proceedings' or 'conduct of a juror in connection with those proceedings may provide grounds for an appeal against conviction or sentence".

Either condition alone is sufficient to justify disclosure. No further definitions are given in the Act for these terms, a shortcoming I will discuss in section IV.B.(iii).

[67] ibid [69-70].

[68] ibid [77].

[69] Judges could sanction jurors for contempt, a feature of the common law, before this legislation, and they still can do so after the Act's passage. The new offences created by the 2015 legislation would be statutory offences triable before a jury in the Crown Court, unlike contempt. Some advocates of the law maintained that this threat of sanction via parliamentary statute, instead of judicial warning, would provide extra power to judges' admonitions to jurors. A.T.H Smith, 'Repositioning the law of contempt: The Criminal Justice and Courts Act 2015' (2015) 11 Crim LR 845, 849.

[70] Criminal Justice and Courts Act 2015, Section 73.

[71] This section additionally supersedes the Contempt of Court Act 1981's prohibition on revealing deliberations, though the prohibition is quite similar—the main change arises in the exceptions.

Jurors are allowed to report these two scenarios to the trial judge, the Court of Appeal, the registrar of criminal appeals, the Criminal Cases Review Commission, or the police. Section 74 also allows jurors to report to a "member of staff of [the trial] court who would reasonably be expected to disclose the information only to a person mentioned in paragraphs (b) to (d)", although the exact interpretation of such a phrase is unclear.

The creation of new offences and the reporting exemptions should not be seen as separate parts of the law, but as an integral whole; the provisions for reporting misconduct were loosened precisely to enable the investigation and prosecution of the misconduct offences the Act creates. The intent of the legislation's jury provisions, as reflected in parliamentary debate, focused on the new prohibitions on juror research and misconduct.[72] Members of Parliament particularly spoke on the new prohibitions against internet research; members variously expressed the purpose of the jury sections as, "modernising the law on the work of juries,"[73] or ensuring "the law on jurors and the use of the internet keeps up to date with the march of technology".[74] Eliminating this type of juror misconduct will likely protect defendants, but the Act's exceptions to jury secrecy should properly be seen as a necessary tool for the Crown to prosecute juror misconduct, rather than as a defendant's rights law. For this, and several other reasons, the Act does not adequately solve the dilemma left open by *Mirza*.

B. The act ultimately proves inadequate

(i) The Act's Reporting Provisions Are Overly Institutional and Not Adversarial

The first major limitation of the Criminal Justice and Courts Act is its overly institutional and prosecutorial reporting structure. When a juror commits misconduct, that misconduct harms both parties in the case. The Crown is harmed by the deprivation of a fair jury, the weakening of jury secrecy, and in the case of misconduct that arises after a verdict, the inability to prosecute the case again in most circumstances.[75] The defendant is harmed because of the deprivation of an

[72] It is worth noting that little of the debate in the House of Commons focused specifically on the jury portion of the Act; the debate over sentencing and parole provisions drew much more attention. What discussion there was of the jury sections was relatively non-controversial. The bill received its first reading in the House of Commons on 5 February 2014, its first reading in the House of Lords on 18 June 2014, and the Royal Assent on 12 February 2015.

[73] HC Deb 24 February 2014, Vol 576, Col 52.

[74] HC Deb 24 February 2014, Vol 576, Col 64.

[75] The Criminal Justice Act 2003 allows for some appeals from perverse acquittals, but generally in the limited cases of particularly severe offences and with the consent of the Director of Public Prosecutions.

impartial jury and fair trial. Yet the Criminal Justice and Courts Act focuses on the harm done to the Crown, not the defendant.

Consider, for example, the means by which misconduct might be reported. The Act states that, after a verdict, a juror might disclose misconduct to the judge (or other court staff), Court of Appeal, registrar of criminal appeals, Criminal Cases Review Commission, or police.[76] The Act then states that the Court of Appeal or registrar of criminal appeals may disclose this information to defence counsel, if they believe it could be a ground for appeal.[77] This procedure, though, places too much power over disclosure into the hands of the judiciary, rather than the defence counsel. The adversarial tradition of the English legal system is intended to place more power into the hands of the parties to the case, rather than with an inquisitorial judge. In Blackstone's formulation:

> "[t]he impartial administration of justice, which secures both our persons and our properties, is the great end of civil society. But if that be entirely entrusted to the magistracy, a select body of men, and those generally selected by the prince or such as enjoy the highest offices in the state, their decisions, in spite of their own natural integrity, will have frequently an involuntary bias towards those of their own rank and dignity: it is not to be expected from human nature, that the few should be always attentive to the interests and good of the many".[78]

The Act's trust in the Court of Appeal or registrar of criminal appeals to disclose information to the defence relies upon consistently trustworthy people occupying those roles—as is doubtless usually the case. But an adversarial system does not rely on the goodwill of those in power. It rests upon giving power to the opposing parties.

The Crown Prosecution Service, in its guidance to prosecutors, does countenance the possibility that a juror might report misconduct in deliberations directly to defence counsel. In these instances, the guidance states that it might not be in the public interest to prosecute that juror, so long as the juror does not

[76] Criminal Justice and Courts Act 2015, Section 74.

[77] ibid.

[78] William Blackstone, *Commentaries on the Laws of* England, Vol. 3. (1753) 379. See, generally, Stephan Landsman, 'Rise of the Contentious Spirit: Adversary Procedure in 18th Century England' (1990) 75(3) Cornell LR 496.

disseminate the information more widely.[79] This prosecutorial exception still falls short, however, because it places decision-making power with the Crown Prosecution Service. Prosecutors, of course, should work in the interests of justice and not solely conviction, but they are also human. Trust in the goodwill and civic-mindedness of prosecutors is not a substitute for legal protection. A stronger law would create a statutory exemption from prosecution for jurors in this circumstance.

Finally, the adversarial system relies upon open examination of the evidence, rather than secretive investigation. To reference Blackstone once more, "this open examination of witnesses *viva voce*, in the presence of all mankind, is much more conducive to the clearing up of truth, than the private and secret examination taken down in writing before an officer".[80] By delineating a set list of acceptable investigators who might look into jury misconduct, the Act contravenes this principle. For the public to have trust in the jury system, the public should be able to see the mechanism by which the system hears and decides upon allegations of bias. The Act should both give jurors more protection to report racial bias directly to defence counsel and for that evidence to them be admitted in open court, for public scrutiny. To allow for this open and adversarial examination, however, the Act must remove not only allow jurors to speak from the jury room but also allow courts to listen to them.

(ii) The Act Only Addresses Juror Speech, Not Courts' Inquiries

The Act's focus on allowing jurors to break the seal of the jury room on some occasions only goes part of the way towards addressing juror misconduct. An offence against the jury — whether juror misconduct or jury tampering by a party to the case or third party — not only demands prosecution in its own right but also damages the safety of the conviction at issue, though. As argued earlier, the 2015 Act is heavily focused on enabling prosecution of jurors for misconduct. Even if a juror is free to report misconduct to the Court of Appeal or others without fear of prosecution, the Act contains no guarantee that the court will admit that testimony. The law leaves a gap here. Moreover, there have been few, if any, test cases under the new statute regarding post-verdict testimony.[81] As such, the obstacle for defendants is that a court might hold that the 'long line of authorities,' as Lord

[79] 'Juror Misconduct Offences' (*Crown Prosecution Service*, 5 July 2019) <https://www.cps.gov.uk/legal-guidance/juror-misconduct-offences> accessed 6 December 2020. The 'public interest' prong is one portion of the two-part test prosecutors use to determine whether to charge an offence. The other prong is whether there is sufficient evidence. 'Code for Crown Prosecutors' (*Crown Prosecution Service*, 26 October 2018) < https://www.cps.gov.uk/publication/code-crown-prosecutors#section4> accessed 10 December 2020.

[80] Blackstone (n 78) [373].

[81] Searches of Westlaw, BAILII, and Lexis returned no cases in the U.K. Supreme Court or Court of Appeal involving the disclosure exception provisions of the 2015 Act.

Justice Kennedy wrote, bar admission of juror testimony.[82] The 2015 Act, given its failure to explicitly address admissibility, would be unlikely to change this feature of the common law.

The text of the 2015 Act, though, does show some intent to grant relief for defendants whose cases involve juror misconduct. The Act allows jurors to disclose misconduct which, "may provide grounds for an appeal against conviction or sentence".[83] This exception shows that Parliament, in drafting the law, did have some concern for the disadvantages that defendants face from the misbehaviour of the jury. This concern for defendants indicates that the law intends to carve out an exemption to the common law prohibition on juror testimony. It would be a self-defeating and nonsensical interpretation of the law to argue that it allows jurors to report misconduct that might be grounds for appeal but prohibits courts from considering it. Hence, another weakness of the Act is that it clearly evinces an intent to soften Mansfield's Rule but fails to explicitly do so in the actual text of the law, leading to a lack of clarity.

(iii) The Act Fails to Explicitly Countenance Racial Bias as Grounds of Appeal

Another of the Act's central failings is that does not explicitly mention racial bias, or indeed bias at all, as a potential ground of appeal. The Act provides for jurors to disclose deliberations in two circumstances: where "an offence or contempt of court has been, or may have been, committed by or in relation to a juror in connection with those proceedings", or where the "conduct of a juror in connection with those proceedings may provide grounds for an appeal against conviction or sentence".[84] Racial bias might or might not fall within these provisions. The first provision would cover racial bias if it were connected to some other offence — for example, if a juror researched the case online and then mentioned news coverage of the case that made racially-biased jokes during deliberations. Of course, this scenario would be covered by the exemption for an 'offence' regardless of whether the Article was biased. The trickier case is determining whether juror bias might be disclosable conduct when a jury offence or contempt of court is not committed in its own right.

In most of the preceding noteworthy cases of juror bias, no separate offence was committed. In *Mirza*, the jury's racially-biased assumptions were made solely based on the proceedings in the courtroom;[85] similarly, the jury in *Sander* would not have breached any of the research or prohibited conduct sections of the 2015 Act, had it been in force at the time.[86] In considering whether a juror could disclose

82 *R v Miah and Akbar* (n 24).
83 Criminal Justice and Courts Act 2015, Section 74.
84 ibid.
85 *Mirza* (n 4).
86 *Sander* (n 37).

biased deliberations in these instances, one must look to the three elements of the second exception made by the Act: "conduct of a juror", "in connection with those proceedings", and "may provide grounds for an appeal against conviction or sentence".[87] Racial bias on the part of a juror would clearly meet the second and third parts of this test. A juror's impartiality during deliberations is clearly connected with the trial proceedings; as *Sander* held, prejudice on the part of the jury is clearly a ground for appeal.

Thus, the outstanding question left by the 2015 Act is whether racial bias on the part of a juror constitutes 'conduct'. In some cases, racially-biased speech is clearly conduct. For example, using "threatening, abusive or insulting words" to "stir up racial hatred" would be an offence under the Public Order Act 1986.[88] A juror inciting other jurors to racial hatred would fall under the ambit of 'conduct'. The bar at which speech becomes criminal conduct, though, would be a rather high one to meet. The forms of bias present in juror deliberations are rarely as explicit as the forms of hatred that the law considers to be criminal conduct. In *Mirza*, for example, a juror, "described an admonition not to attach importance to the use of an interpreter as 'playing the race card'".[89] This sort of prejudiced view is not an incitement to hatred, but nonetheless calls into question the jury's impartiality. Such a view is more likely than explicit racial animus to arise in a jury room. Would this type of prejudice be 'conduct' for the purposes of the Criminal Justice and Courts Act?

Very few cases amongst the higher courts, if any, have considered the provisions of the Criminal Justice Act that create new offences under Section 20 of the Juries Act 1974. The scant record that exists suggests that the courts are not willing to interpret juror bias as a form of prohibited conduct regarding which deliberations might be disclosed. In one of the cases regarding whether a judge could inquire into potential juror bias, *R v Eaton*, the Court of Appeal held that, "the judge could not have asked the juror what she might have discussed with her fellow jury members by reason of the provision of Section 20 of the Juries Act 1974, as amended".[90] While the bias at issue in *Eaton* was a personal connection to a co-defendant, not racial bias, this case indicates that juror prejudice would likely

87 Criminal Justice and Courts Act 2015, Section 74.
88 Public Order Act 1986, Section 18.
89 *Mirza* (n 4) [28].
90 *R v Eaton* [2020] EWCA 595 [22].

not be considered 'conduct' within the meaning of the 2015 Act. This omission is a weakness of the law.

(iv) The Act's Reporting Procedures Will Discourage Reporting

The Act's vague reporting provisions and exemptions, along with strict criminal penalties for illegal disclosure, will discourage reporting of misconduct and bias by lay jurors. Few jurors will ponder the legal intricacies of whether bias is 'conduct', as discussed in the previous section. Consider, too, the reporting provisions of the bill. Jurors may report misconduct in deliberations to several people, one of whom is "a member of staff of that court who would reasonably be expected to disclose the information only to [the trial judge, Court of Appeal, registrar of criminal appeals, or police]". How would a lay juror interpret such a provision and identify such a staff member? Would a court usher, with whom jurors interact quite often, be 'reasonably expected' to meet this provision? Would a clerk or security officer at the court entrance? Now, juxtapose this confusing process with the Act's overall emphasis on juror punishment for misconduct.[91] The result is that jurors will err on the side of caution and be hesitant to report bias of misconduct.

Indeed, in the face of confusion, jurors usually default to not reporting misconduct at all. A recent study asked jurors at the Old Bailey what they understood the restrictions of the 2015 Act to be, after their receiving a notice about it; just under three-quarters of jurors correctly identified the restrictions placed on them.[92] More interestingly, of the jurors who did *not* correctly identify the rules about jury offences, the majority interpreted the rules too strictly — that is, they believed there were no exceptions to the prohibition of disclosure.[93] This study indicates that, in the face of complex disclosure procedures and potential criminal penalties, jurors will likely be overly cautious and not report abnormalities.

Members of Parliament critiqued the bill as unclear during debate in the House of Common. Then-MP Sadiq Khan noted that jurors needed further education and training to correctly follow the law, saying,

> "[t]here are problems with juries not understanding their role sufficiently, and we shall explore what steps can be taken to educate and inform the public and jurors about the important civic function of jury service so that it is less of an alien process to them".[94]

[91] See generally Kevin Crosby, 'Juror Punishment, Juror Guidance, and the Criminal Justice and Courts Act 2015' (2015) 8 Crim LR 578.

[92] Cheryl Thomas, 'The 21st Century Jury: Contempt, Bias and the Impact of Jury Service' (2020) 11 Crim LR 987, 996.

[93] ibid.

[94] HC Deb 24 February 2014, Vol 576, Col 64.

MP Andy Slaughter criticised the government for not ensuring that jurors would be provided this training, especially in the digital age, saying, "[h]owever the Government fail to provide any support to juries in explaining their roles and remit as part of any new offences […]".[95] The Act's lack of clarity around reporting is a major stumbling block.

In summary, the Criminal Justice and Courts Act is a meaningful but ultimately too limited step towards jury transparency. Its reporting procedures suffer from several defects — opacity, the failure to explicitly recognise racial bias, and a lack of adversarialism — that make them unlikely to meaningfully combat racial bias in the jury room. Moreover, the Act only allows jurors to speak; it does not allow courts to listen. Such a remedy will always be incomplete, especially given how much weight courts have traditionally given to the common-law Mansfield's Rule. Better reporting procedures and an explicit statutory repeal of Mansfield's rule are needed.

V. Reforming the 2015 Act

A. Expanding criminal liability exception to racial bias

Parliament should amend the Criminal Justice and Courts Act to specifically allow jurors to disclose deliberations in the event of racial bias on the part of a juror or jurors. As argued in section IV.B.(iii), it is unlikely that racial bias is covered by the current Act's exemption for 'conduct'; Parliament should add to the statute an explicit statement that jurors may disclose racial bias from the deliberations. *Mirza*, and Lord Steyn's dissent in particular, clearly identified this lacuna in English law; when grievous racial bias threatens an individual defendant's right to a fair trial, the broader interest of jury secrecy takes on comparatively less importance. The Criminal Justice and Courts Act worked part way towards remedying this problem, but fell short for a number of reasons, as argued in the previous section. Racial bias is, as the ECtHR in *Sander* held, a unique threat imperilling the right to an impartial tribunal. Its specific inclusion in the statutory exemptions to jury secrecy would help guarantee more defendants the right to an impartial jury. This exemption would also improve the legitimacy of the jury system, especially amongst minority racial and ethnic groups, by showing that the government recognises and is combatting the threat of bias.

In crafting such an amendment, one should first consider what the standard for such bias should be. The Court of Appeal held in *In Re Medicaments* that the European Convention on Human Rights requires a tribunal to assess whether the circumstances regarding a judge or juror, "lead a fair-minded and informed observer to conclude that there was a real possibility, or a real danger, the two

[95] HC Deb 24 February 2014, Vol 576, Col 121.

being the same, that the tribunal was biased".[96] This stipulation that bias must actually be expressed and cast doubt on the conviction's fairness would help to protect against frivolous accusations, one of the concerns of the majority of the Lords in *Mirza*.[97] Jurors must not report bias simply based on a general impression, but rather should report particular remarks or events in the jury room that create the 'real possibility' of bias. Parliament, however, ought to emphasise the ability of lay jurors to understand the statute. While judges should instruct jurors to report actual events or remarks that occur, not feelings or intuitions, the consideration of whether bias was actually pernicious enough to affect the conviction ought to be a judicial matter. The law should provide jurors the broadest exception from prosecution possible for good faith reporting. Such a construction would give jurors more confidence to report racial bias; the court could then consider whether or not the conviction should stand. The courts could maintain a rebuttable presumption that the jury was fair, as the law already dictates, in the interests of finality.[98] Hence, statutory language such as the following might suffice: "It will not be an offence to disclose a juror's comment or remarks during deliberations that might indicate racial prejudice or bias towards the defendant". Of course, to enable this judicial consideration, Parliament must also allow courts to admit juror testimony — section V.B addresses this issue.

The other crucial question is whether this exception should allow jurors to report solely racial bias, or also biases against other identity characteristics.[99] Racial bias is not the only type of bias that Parliament might allow jurors to disclose from deliberations, but it should be prioritised. In *Mirza*, Lord Hope raises this question of whether racial bias is deserving of different treatment than discrimination based on language, social group, religion or other factors.[100] Something must distinguish racial bias from these other biases if it is to merit a special exception to jury secrecy rules. Crafting exceptions to jury secrecy is a balancing act. There are good reasons to not want every conviction thrown into doubt by post hoc juror testimony; jury secrecy does serve a legitimate purpose in fostering frank and open deliberations. Understandably, Parliament might not want to allow exceptions to

[96] In *Re Medicaments* [2000] EWCA Civ 350, at para. 85. This standard, though applied to the judge in In Re Medicaments, applies equally to juries as well. R v Brown [2001] EWCA Crim 2828.

[97] *Mirza* (n 4) [53–54].

[98] ibid [112].

[99] Even within the umbrella of 'racial discrimination', the Equality Act 2010 defines race so as to include ethnicity and national origin, colour, and nationality.

[100] *R v Mirza* (n 4) [77].

juror secrecy for every possible type of bias jurors detect from their fellow jurors during deliberations.

Race, however, has several characteristics which make it particularly important; race is foremost amongst those characteristics which merit special exceptions for the reporting of bias. First, as argued earlier, the systemic overrepresentation of black people amongst those stopped by police, those prosecuted by the CPS, and those sentenced to prison shows that race matters in criminal justice.[101] This prima facie evidence of racial disparities gives race importance. Second, minority racial and ethnic groups have faced and continue to face discrimination in other realms such as housing and healthcare.[102] Racial bias clearly affects other areas of society and has a likelihood of being present in jury deliberations. Indeed, in a U.S. case analogous to *Mirza*, *Peña-Rodriguez v Colorado*, the U.S. Supreme Court held that racial bias was such a threat that evidence of it in jury deliberations should be admitted in post-conviction challenges, contravening the traditional application of Mansfield's Rule.[103] As the U.S. Supreme Court noted, racial bias is, "a familiar and recurring evil that, if left unaddressed, would risk systemic injury to the administration of justice".[104] Moreover, as detailed in section III.B.(ii), other safeguards are unable to ensure an unbiased jury with regard to race. Other forms of discrimination are not unimportant, but racial bias, in the criminal justice context, is especially pernicious. Based on the 2015 Act, Parliament does not want to open to floodgates to massive exceptions to jury secrecy. Racial bias should be prioritised, though, amongst a limited number of exemptions as part of this balancing act.

Parliament should additionally change the procedure by which jurors can report bias under the Criminal Justice and Courts Act. As argued in section III.B.(i) and III.B.(iv), the procedures under the 2015 Act are both insufficiently adversarial and overly complicated. They therefore deserve change in their own right to effectively implement the very purpose of the Act, which is to respond to juror misconduct. Hence, I propose two changes. First, defence attorneys should be added as a safe party to whom jurors may disclose misconduct without fear

[101] Lammy Review (n 40).

[102] Cabinet Office, 'Race Disparity Audit' (*GOV.UK*, October 2017) <https://www.gov.uk/government/publications/race-disparity-audit> accessed 9 December 2020 and Public Health England, 'Beyond the data: Understanding the impact of COVID-19 on BAME groups' (June 2020) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/892376/COVID_stakeholder_engagement_synthesis_beyond_the_data.pdf> accessed 9 December 2020.

[103] *Peña-Rodriguez v Colorado* 580 US (2017).

[104] ibid 16.

of prosecution. This change would recognise the adversarial nature of the legal system and empower the defendant's legal representatives.

Additionally, Parliament should make the CPS charging guidance, which generally advises against charging jurors who report misconduct to involved parties in good faith, into law. This guidance states that it is not in the public interest to charge jurors who report misconduct to appropriate parties, for the purpose of assisting the court to hear the case fairly, rather than spreading it to the media or public.[105] It is unlikely that jurors will have consulted the CPS guidance and know this, though. To provide jurors more reassurance, and to thus increase reporting of misconduct to the court and involved parties, Parliament should make this guidance into law, and the courts should incorporate it into the notice given to all jurors. At the same time, courts might continue to reinforce that leaking jury deliberations to the media or on the internet would still result in criminal penalties. This change would encourage jurors to report misconduct while maintaining many important benefits of jury secrecy. It would also remove discretion from prosecutors — who have a vested interest in not seeing their convictions disturbed — over whether to charge jurors who report misconduct. Jurors might report misconduct and bias more readily if they know they will not face prosecution when they are earnestly trying to do right by the court and defendant.

B. Changing mansfield's rule

Racial bias is a pernicious threat to defendants' rights and the legitimacy of the criminal legal system that jurors should be able to speak about post-verdict. If one accepts the contention that the bar on jurors speaking about deliberation should be removed, there are no justifiable grounds to not also make an exception to Mansfield's Rule and allow courts to consider this information to quash a conviction. If the intent of the 2015 Act is at least partially to allow defendants some remedy for jury misconduct, as argued in section IV.B.(ii), then courts must be able to consider that misconduct. The courts are the sole body able to grant relief to a defendant whose conviction is unsafe because of juror racial bias;[106] to have any effect, reform of jury secrecy must allow for juror evidence to be considered by the court.

There are, of course, legitimate interests behind Mansfield's Rule — finality and protecting the candour of deliberations. Both of these interests are already rendered moot, though: Once jurors are free to speak about misconduct, the finality of the conviction is undermined in the public eye. Once information

[105] 'Juror Misconduct Offences' (n 79).

[106] The Criminal Cases Review Commission, the other major body involved in reviewing the safety of convictions, may only refer cases back to the Court of Appeal.

from deliberations is released, future jurors' candour is, in theory, chilled. It would be a graver injustice to allow convictions to stand solely on the formalism that finality must be upheld, especially because the benefits of secrecy were obviated by removing the gag on jurors. The courts should use the objective standard from *In re Medicaments* to assess whether a conviction was tainted by bias and potentially grant relief.[107] Wilful ignorance towards bias helps no one, delegitimises the legal system, and makes it look hapless in the face of injustice.

Removing Mansfield's Rule is not a cure-all for every type of bias; courts might still be reluctant to consider allegations of implicit, rather than explicit, racial bias. This type of bias is often harder to recognise and harder to prove following a conviction. While there are additional measures courts might take to fight the problem of implicit bias, this limitation is no good reason to remain ignorant of explicit bias in the meantime.[108] Parliament must amend the Act to directly state that evidence of juror misconduct and bias may be admitted by the courts in considering a defendant's appeal.

VI. Conclusion

The disproportionate representation and disparate treatment of people from marginalised racial and ethnic groups at each stage of the criminal justice system raises grave concerns about where racial bias might arise. Moreover, to maintain the public legitimacy of the jury system, people must have faith that it is impartial and fair; given the higher proportion of Black and Asian defendants that opt for a jury trial, this legitimacy with regard to racial bias is even more paramount. Other safeguards, such as oaths and random selection, might help in preventing some bias, but they are of no use once bias or prejudice occurs; then, only judicial review of the conviction can help the defendant. When jurors cannot speak out and courts cannot consider the evidence, the jury system cannot maintain its own fairness. *Mirza* highlights this approach to bias, where the interests of finality and secrecy overpower the interests of fairness and legitimacy.

The Criminal Justice and Courts Act 2015 partially helps in unsealing the jury room but does not extend far enough. Besides not explicitly combatting racial bias, the Act's reporting provisions are complex and unclear. Combined with its harsh criminal penalties, the Act might ultimately discourage juror reports of misconduct. Finally, the Act does nothing to undo the common law rule against admitting juror testimony. But there is a possible solution. Reforming the Criminal Justice and Courts Act based on the lessons of *Mirza* would be one

[107] In *Re Medicaments* (n 9).

[108] For an overview of potential measures to combat implicit bias on juries, see Anona Su, 'A Proposal to Properly Address Implicit Bias in the Jury' (2020) 31 Hastings Women's Law Journal 79.

small step toward fairness. Parliament should allow jurors to report, without fear of prosecution, instances of racial bias from deliberations to the court or the defendant's legal counsel. Furthermore, the law should allow courts to consider the testimony of jurors in quashing convictions because of the jury's racial bias or other misconduct. The legal system's insistence on ignoring bias in the jury room does not make that bias disappear. It does not make the tainted convictions any fairer. Instead, this wilful refusal allows prejudice to erode the foundations of impartiality upon which the jury system rests. If the jury is to survive as a legitimate feature of English criminal justice, Parliament must allow the courts to look inside the jury room when prejudice, discrimination, and bias lurk. What happens in the jury room must not stay there.

Cambridge Law Review (2021) Vol VI, Issue i, 263–290

Spandeck: A Relational View of the Duty of Care

Soh Kian Peng*

Abstract

The use of a general framework in the determination of a duty of care has seemingly fallen out of favour following the UK Supreme Court's decision in *Robinson*. Relying on the example of the *Spandeck* framework in Singaporean jurisprudence, this piece presents the argument that such frameworks, being consistent with a relational conception of tort law, can provide a useful means of determining whether a duty of care exists. In so doing, this piece addresses some criticisms of the relational view and re-emphasises the important role the duty of care plays in the tort of negligence.

Keywords: tort, negligence, corrective justice, duty of care, Spandeck

I. Introduction

Almost 90 years have passed since the seminal judgment in *Donoghue v Stevenson*.[1] Yet the duty of care concept remains fraught and contested.[2] The lack of a clear approach is problematic.[3] Tort law, being a "social and evolutionary phenomenon

* LL.B., summa cum laude (Singapore Management University). I would like to thank Professor Gary Chan for his comments on earlier drafts of this article, and Associate Professor Maartje De Visser, Mr Vincent Ooi and Ms Ong Ee Ing for their encouragement. The usual caveat applies. kianpengsoh.2017@law.smu.edu.sg.

1 [1932] AC 562.

2 See *Robinson v Chief Constable of West Yorkshire Police* [2018] 2 All ER 1041 [21], [30], [83], [100].

3 Andrew Clarke and John Devereux, 'Hard Cases Making Bad Law: The Elusive Search for a Test for Duty of Care' (2019) 26 Tort L Rev 177, 183.

[…] where the law and social life affect each other in complex ways"[4] must therefore continually adjust to rapidly changing social circumstances. Courts may soon be invited to decide whether duties of care exist in novel cases.[5] While existing legal principles may be extended to cover unique factual matrices that may arise,[6] these legal principles must be coherent if they are to be meaningfully applied.

This article therefore argues that the general framework set out by the Court of Appeal in *Spandeck Engineering (S) Pte Ltd v Defence Science & Technology Agency* represents a clear and principled approach to analysing a duty of care.[7] Writers have expounded on the merits of the *Spandeck* framework[8] or have attempted to flesh out the concept of proximity,[9] but I aim to add to this literature by showing how *Spandeck* is consistent with a relational theory of tort law.[10]

Following the introduction in Part I, this article proceeds in four parts. Part II lays out the features of the *Spandeck* framework. Part III sketches out how a relational theory of tort is reflected in the *Spandeck* framework. Criticisms of the relational view and of proximity will be addressed, along with some implications arising from the relational view of tort. Part IV explains, with reference to cases, how *Spandeck* reflects this relational view sketched out in Part III. Slight changes are proposed to the *Spandeck* formulation to better align it with the relational view. Part V concludes.

II. Features Of *Spandeck*

Spandeck is a two-stage test prefaced by the threshold requirement of factual foreseeability. The threshold requirement of factual foreseeability is a low one that will invariably be satisfied in most cases.[11] Here, the courts examine the facts to

4 Peter Cane, *Key Ideas in Tort Law* (Bloomsbury, 2017) 81–82; Goh Yihan, 'Tort Law in the Face of Land Scarcity in Singapore' (2009) 26(2) Arizona J of Intl & Comparative L 335.

5 *Oscar Willhelm Nilsson v General Motors LLC* (ND Cal) (Trial Pleading) WL 514625 (2018). The Plaintiff in this case was involved in an accident with a self-driving vehicle. He sued General Motors ("GM") in the tort of negligence, alleging that GM owed him a duty to have its self-driving vehicle operate in a manner which obeyed traffic laws and regulations.

6 *Dorset Yacht Co. Ltd v Home Office* [1970] AC 1004, 1026–27.

7 [2007] 4 SLR(R) 100 [72]. See also David Tan and Goh Yihan, 'The Promise of Universality' (2013) 25 SAcLJ 510 [4]–[8]; *Toh Siew Kee v Ho Ah Lam Ferrocement (Pte) Ltd and others (CA)* [2013] 3 SLR 284 [54].

8 David Tan, 'The End of the Search for a Universal Touchstone for Duty of Care?' (2019) 135 LQR 200.

9 David Tan, 'The Salient Features of Proximity: Examining the *Spandeck* Formulation for Establishing a Duty of Care' (2010) SJLS 459, 469 – 481.

10 See generally Ernest Weinrib, 'The Disintegration of Duty' (2006) 31(2) Advocates Quarterly 212, 233–45.

11 *Spandeck* (n 7) [75]–[76].

determine if it would have been foreseeable to the defendant that a failure to take reasonable care would result in the plaintiff suffering loss.[12]

At the first stage, the court considers whether there is legal proximity between the parties.[13] Proximity includes "physical, circumstantial and causal proximity" and the "twin criteria of voluntary assumption of responsibility and reliance" ('VAR-R'),[14] and has been expanded to include other factors, such as knowledge.[15] If the proximity requirement is met, a *prima facie* duty of care arises.[16] At the second stage, policy factors militating against the imposition of this duty are considered. This involves a "weighing and balancing of competing moral claims and broader social welfare goals".[17] Examples of policy factors include the existence of a contractual framework,[18] indeterminate liability,[19] and the value of human life.[20] Policy reasons that favour imposing a duty of care can be considered to dismiss the defendant's "spurious negative policy considerations".[21]

III. Clearing the Conceptual Ground

Proximity is central to the *Spandeck* framework.[22] Other jurisdictions, however, have utilised concepts such as "reasonable foreseeability"[23] or policy reasons in the duty of care analysis.[24] Here I address some criticisms of proximity in the duty of care analysis, arguing that this analysis is best approached through the concept of proximity because it reflects the essence of tort law which is, on the relational view, primarily concerned with corrective justice.[25]

A. Addressing criticisms of proximity

There are two main criticisms against using proximity to determine the existence of a duty of care: first, proximity merely expresses "the result of

12 ibid [89]; *Animal Concerns Research & Education Society v Tan Boon Kiwee* [2011] 2 SLR 146 [35].
13 *Spandeck* (n 7) [77]–[82].
14 ibid [81].
15 *NTUC Foodfare Co-operative Ltd v SIA Engineering Co Ltd and another* [2018] 2 SLR 588 [50].
16 *Spandeck* (n 7) [83].
17 ibid [85].
18 ibid [114].
19 *NTUC Foodfare* (n 15) [54].
20 *Man Mohan Singh s/o Jothirambal Singh and another v Zurich Insurance (Singapore) Pte Ltd* [2008] 3 SLR(R) 735 [51]; *ACB v Thomson Medical Pte Ltd and others* [2017] 1 SLR 918 [210].
21 *Animal Concerns* (n 13) [77].
22 *Spandeck* (n 7) [79]–[81].
23 See Stephen Todd (ed), *The Law of Torts in New Zealand* (6th edn, Thomson Reuters 2013) [5.2.03].
24 See *Robinson* (n 2) [29], [30], [42]; *Caltex Refineries (Qld) Pty Ltd v Stavar* (2009) 75 NSWLR 649.
25 See John Gardner, 'What is Tort Law For? Part 1. The Place of Corrective Justice' (2011) 30(1) Law and Philosophy 1, 6.

a process of reasoning rather than the process itself";[26] and, second, proximity has been described as a mere label in contrast with a proper concept insofar as a duty of care is concerned.[27] Plunkett, for example, cites Mason CJ and Wilson J's dissent in *Hawkins v Clayton*,[28] arguing that proximity is a mere label and pointless as a concept.[29] Both judges opined that the relevant inquiry was whether "the professional relationship of solicitor and client gave rise to a *relationship of sufficient proximity* founded upon an assumption of responsibility […] and reliance".[30] Plunkett argues that the reference to "more specific concepts"[31] in determining the existence of a duty supports the aforementioned criticisms of proximity. A closer examination of the judgement, however, suggests that both judges used proximity *qua* descriptor and not *qua* concept. Both judges concluded that "intermeddling in the estate has no bearing on the existence or otherwise of the *requisite relationship of proximity* […] sufficient to found the alleged duty".[32] Clearly, both judges expressed the result of their analysis by saying that there was no "relationship of proximity".[33]

Criticisms of proximity therefore stem from the lack of a clear understanding of the context in which 'proximity' is used.[34] Where there is a duty of care, the parties are in sufficient *proximity* to each other. We express the results of our analysis accordingly: "a duty arises because the parties are sufficiently proximate" or there was a "relationship of proximity". In these statements, proximity expresses the result of finding that there is a duty of care in a particular situation. But that is different from the idea of proximity *qua* concept.

Moreover, 'proximity' in common parlance gives the impression of the parties being close in space and time.[35] One might interpret the statement "a duty arises because the parties are sufficiently proximate" to mean that a duty arises because both parties are sufficiently close to each other in *time* and *space* such that

[26] *Hill v Van Erp* (1997) 71 ALJR 487, 558. See also James Plunkett, *The Duty of Care in Negligence* (Hart Publishing 2018) 188.

[27] *Caparo v Dickman* [1990] 2 AC 605, 628. Cf Andrew Phang, Cheng Lim Saw, and Gary Chan, 'Of Precedent, Theory and Practice - The Case for a Return to Anns' (2006) SJLS 1, 41–42.

[28] (1988) 78 ALR 69.

[29] Plunkett (n 26) 188.

[30] *Hawkins v Clayton* (1988) 78 ALR 69, 72 (emphasis added).

[31] Plunkett (n 26) 188.

[32] ibid 73 (emphasis added).

[33] ibid.

[34] See David Adger, 'This Simple Structure Unites All Human Languages' (2019) 76 Nautilus <http://nautil.us/issue/76/language/this-simple-structure-unites-all-human-languages> accessed 27 September 2019.

[35] See Low Kee Yang, 'Occupiers' Liability After See Toh: Change, Uncertainty and Complexity' (2013) SJLS 457, 468.

one party ought to take reasonable care, by bearing in mind the other party, when acting.[36]

However, the definition of "proximity" extends *beyond* temporal and spatial relationships. For instance, in *Spandeck* the court relied heavily on the *Sutherland* factors[37] which not only include physical and causal, but also circumstantial, proximity which Deane J in *Sutherland* described as "an overriding relationship of employer and employee".[38] Subsequent cases, applying the concepts of VAR-R or knowledge to establish the presence of a *prima facie* duty of care, have expanded the scope of proximity beyond the temporal and spatial aspects. In cases relying on VAR-R, it would be a stretch to use proximity in terms of being close in time and space. Proximity in these cases demonstrates a different meaning: that both parties are close in terms of *moral* relationships.[39]

Courts are aware of the propensity of language to confuse. In *NTUC Foodfare Co-operative Ltd v SIA Engineering Company Limited* ('*NTUC Foodfare*'),[40] the case involved a claim for pure economic loss arising from the defendant's negligent operation of an airtug which crashed into a pillar. This caused structural damage, affecting the plaintiff's food kiosk which was situated nearby. Consequently, the plaintiff was forced to shut its food kiosk. The court held that there was sufficient legal proximity between the plaintiff and the defendant to found a duty of care.[41] This was due to "physical proximity between the parties" as the defendant was "operat[ing] airtugs in close *propinquity* to the [plaintiff's] [k]iosk".[42] Using *propinquity* instead of *proximity* signifies that the court did not want to confuse proximity *qua* legal concept and proximity *qua* descriptor in describing the facts.

Clearly, the context in which "proximity" is used distorts its meaning *qua* concept and meaning *qua* descriptor. This confusion, however, can be resolved by understanding that proximity refers to a set of intrinsic characteristics and its centrality in the duty of care analysis which the court in *Spandeck* alluded to in that "proximity has some substantive content that can be expressed in terms of legal principles".[43] The linguistic meaning of proximity in this context is that of

[36] See also Justin Tan, 'Proximity as Reasonable Expectations' (2019) SJLS 147, 167.
[37] Spandeck (n 7) [81] citing *Sutherland Shire Council v Heyman* (1985) 60 ALR 1.
[38] *Sutherland Shire Council v Heyman* (1985) 60 ALR 1, 55–56.
[39] See Section C "Proximity Defined qua Concept" below.
[40] [2018] 2 SLR 588.
[41] ibid [46].
[42] ibid [47] (emphasis added).
[43] Spandeck (n 7) [80]. See also *Turf Club Auto Emporium Pte Ltd v Yeo Boong Hua* [2018] 2 SLR 655 [183]–[185] where the Singapore Court of Appeal distinguished between descriptive and normative restitution. The former does not shed "any light on why the gains were disgorged as well as the conceptual basis of the relevant head of damages".

a *concept* bearing certain *essential* characteristics,[44] not of a *descriptor*. Having made this crucial distinction, the following sections flesh out what proximity *qua* concept means and how it instantiates a relational view of tort law.

B. Defining concepts

There are three possible ways of defining proximity *qua* concept. First, through essentialism, concepts are defined by drawing from the essence of the concept itself.[45] It arises from the idea that everything has a basic set of characteristics. The process of defining involves "isolating this common nature or intrinsic property".[46] Second, concepts may also be defined through linguistic use: the definition of the concept arises from the manner of its linguistic usage.[47] The traditional interpretation of Wittgenstein's 'family resemblance' passages is a straightforward denial of essentialism: there is no essentialist definition that captures the common features of a concept-word.[48] Concept-words therefore only derive their identity from "a shareable practice of expression, reaction and use of language".[49] Bangu's alternative interpretation of Wittgenstein posits that "speakers do not need to know an essentialist definition of games in order to apply [the word] game[s] correctly".[50] One does not "feel the pressure of the requirement to be able to identify a common feature while we use the terms correctly".[51] For instance, one does not need to identify common features of games to use the word 'game' correctly.

Where essentialism is concerned, concepts clearly do not exist independently of language. On the other hand, Bangu might have a point that everyday users of language need not know the common features encapsulated by a word to use that word correctly. However, where the law is concerned, and concept-words are used to denote or refer to certain ideas, one must know what these ideas are to correctly use the concept-word. For example, to use 'consideration' in contract law correctly,

[44] See Desmond Manderson, 'Emmanuel Levinas and the Philosophy of Negligence' (2006) 14 Tort L Rev 33, 46.

[45] Michael Freeman (ed), *Lloyd's Introduction to Jurisprudence* (9th edn, Sweet & Maxwell 2014) [1-009].

[46] ibid.

[47] ibid [1-008].

[48] Sorin Bangu, 'Later Wittgenstein on Essentialism, Family Resemblance and Philosophical Method' (2005) 6(2) Metaphysica 53, 56.

[49] Stewart Candlish and George Wrisley, 'Private Language' (Stanford Encyclopaedia of Philosophy, 30 July 2019) <https://plato.stanford.edu/entries/private-language/#SigIss> accessed 27 September 2019.

[50] Sorin Bangu, 'Later Wittgenstein on Essentialism, Family Resemblance and Philosophical Method' (n 48) 62.

[51] ibid.

one must know the bundle of ideas (i.e., an element in the formation of a valid contract) to which it refers.

The third way of defining concepts, termed by Zipursky, is "pragmatic conceptualism".[52] In accordance with this view, concepts are understood by grasping from "within the practices of the law, the pattern of verbal and practical inferences that constitute the relevant area of the law".[53] While the starting point of any concept focusses on linguistic expression,[54] and concepts can be "partially shaped by linguistic practices, this does not necessarily entail that concepts are meanings".[55] For instance, considering the various *concepts* of law, *viz.*, law as a series of general orders backed up by threats (Austin) or law as a union of primary and secondary rules (Hart), this differs from how lawyers or laypeople use the word 'law'. Per Canale, "conceptual content does not identify with linguistic content, although the former is strictly related to the latter".[56]

Pragmatic conceptualism holds that the rules and principles of tort, which are not *identical* to their verbal formulations, can be found in the *practice* of participants of the legal community.[57] While linguistic usage of proximity can confuse, a closer look at how the Singapore courts have used "proximity" in the *context* of the *Spandeck* framework suggests that it refers to certain principles of tort law.

Next, I sketch out how 'proximity' and the *Spandeck* framework are intrinsically tied to a *relational* view of tort law, and I address critiques of the relational view, arguing that it can accommodate instrumental concerns present in policy reasoning.

C. Proximity defined *qua* concept

(i) The Relational Theory of Tort and of Proximity

A perusal of cases demonstrates that the Singapore courts have used proximity to denote the existence of a "relationship between the tortfeasor and

[52] Benjamin C Zipursky, 'Pragmatic Conceptualism' (2000) 6 Legal Theory 457.
[53] ibid 473.
[54] Damiano Canale, 'Consequences of Pragmatic Conceptualism: On the Methodology Problem in Jurisprudence' (2009) 22(2) Ratio Juris 171, 173–74.
[55] ibid.
[56] ibid.
[57] Hanoch Dagan and Benjamin Zipursky, 'The Distinction between Private Law and Public Law' 18 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3641950> accessed 27 September 2019.

the claimant insofar as it is relevant"[58] to the loss suffered by the claimant. Factors such as causal, physical and circumstantial proximity indicate when proximity is made out in a particular case.[59] Therefore, the *Spandeck* framework deals with the fundamental question of whether a relationship exists between the tortfeasor and claimant in the present case such that the law of negligence should apply.[60]

As I seek to sketch out in this section, and explain in Part IV, this is consistent with a relational, as opposed to an instrumentalist, view of tort law which conceives of tort law as a mechanism for pursuing "collective goals such as economic efficiency and loss spreading".[61] The problem with the instrumentalist view is that in explaining the *function* of tort law, it glosses over the importance of the concept of a duty of care.[62]

Relational theories, however, conceptualise tort law as regulating "certain kinds of interactions or transactions between" [63] people. A duty of care is owed if our behaviour would result in "some other aspect of a person's life being damage[d] or imperilled".[64] This is expressed in terms such as 'interactional', 'transactional', 'bipolar', 'bilateral' and 'correlative'.[65] A duty, if breached, gives rise to a corresponding right *in personam*.[66] Significantly, the relational view reveals the moral aspect of tort law based on corrective justice. Tort law is concerned with the relationship between "the defendant's doing and the plaintiff's suffering"[67] harm as a consequence.

That duties cannot be owed to strangers is one objection to the relational view.[68] Howarth observes that "defendants have had no relationship[s] at all with

[58] *Go Dante Yap v Bank Austria Creditanstalt AG* [2011] 4 SLR 559 [32]; see *Toh Siew Kee* (n 7) [53]; *Jurong Primewide Pte Ltd v Moh Seng Cranes Pte Ltd and others* [2014] 2 SLR 360 [37]; *Ramesh s/o Krishnan v AXA Life Insurance Singapore Pte Ltd* [2015] 4 SLR 1 [243] ('Ramesh'); *NTUC Foodfare* (n 15) [43]; *Minichit Bunhom v Jazali bin Kastari and another* [2018] SGCA 22 [2].

[59] Norman Katter, 'Who Then in Law is My Neighbour? Reverting to First Principles in the High Court of Australia' (2004) 12 Tort L Rev 85, 97.

[60] See *Caparo v Dickman* [1990] 2 AC 605, 363.

[61] Stephen Perry, 'Torts, Rights, and Risk' in John Oberdiek (ed), *Philosophical Foundations of the Law of Torts* (OUP 2018) 39.

[62] ibid 41. See Robert Stevens, *Torts and Rights* (OUP 2012) 2; Canale, (n 54) 484; Kenneth M Ehrenberg, *The Functions of Law* (OUP 2016) 5.

[63] Stephen Perry, 'Torts, Rights and Risk' (n 61) 38–64.

[64] John Gardner, *From Personal Life to Private Law* (OUP 2018) 50.

[65] Stephen Perry, 'Torts, Rights and Risk' (n 61) 40.

[66] Wesley Newcomb Hohfeld, 'Some Fundamental Legal Conceptions as Applied in Judicial Reasoning' (1913) 23 Yale LJ 16; Canale (n 54) 463.

[67] Ernest Weinrib, 'The Special Morality of Tort Law' (1989) 34(3) McGill LJ 403, 408; John Gardner, *From Personal Life to Private Law* (n 64) 50.

[68] Nicholas J McBride, 'Duties of Care - Do They Really Exist?' (2004) 24(3) OJLS 417, 433.

their claimants".[69] Even if we can owe a duty of care to strangers, we cannot give reasons for owing such duties. Reasoning that a duty of care arises out of some relationship with a potential victim entails a perverse "view of what counts as a relationship".[70] Howarth thus concludes that some "tort duties derive from general law" and not from relations or relationships.[71] According to Howarth, where strangers are concerned, the relationship between the wrongdoer and sufferer only crystallises at the point the tort is occasioned.[72] Therefore, because no such relation existed prior to the commission of the tort, the relational view cannot explain why we owe duties of care to strangers.

However, counterfactuals can explain why a duty of care exists on a relational view.[73] Say, for example, I injure a pedestrian because of my negligent driving. *Pace* Howarth, absent a relationship between me and the victim when the tort was committed, the relational view cannot explain why a duty is owed.[74] Counterfactuals, the use of which is not alien to tort law (i.e., the "but-for" test in causation,[75] and the assessment of damages[76]), can explain this. In the counterfactual, we can imagine the identical situation of driving along the same road, except that no accident took place this time. With knowledge of the facts that an accident that resulted in injury *did* occur, one can ask whether a relationship should exist between the potential wrongdoer and sufferer such as to impose a duty of care on the potential wrongdoer. This is known as "conceptual blending".[77]

[69] David Howarth, 'Many Duties of Care - Or a Duty of Care? Notes from the Underground' (2006) 26(3) OJLS 449, 463.

[70] ibid 464.

[71] ibid.

[72] ibid 463–64.

[73] See Steven L Winter, 'Frame Semantics and the Internal Point of View'. in Michael Freeman and Fiona Smith (eds), *Law and Language: Current Legal Issues* Vol 15 (OUP 2013).

[74] See *Cameron v Liverpool Victoria Insurance Co Ltd* [2019] UKSC 6 (holding that procedural rules bar a claimant from suing a totally anonymous person).

[75] Michael Jones (ed), *Clerk and Lindsell on Torts* (20th edn, Sweet and Maxwell 2010) [2-09].

[76] ibid [28-07].

[77] Steven L Winter, 'Frame Semantics and the Internal Point of View' (n 73) 120.

It involves projecting oneself into an alternate mental space whilst retaining knowledge of the facts at hand in a manner described above.[78]

Our use of counterfactuals reveals deeper implications,[79] capturing our view of moral responsibility.[80] Our ability to empathise enables us to consider the counterfactual.[81] It demonstrates that we are not *merely* neighbours in a "temporal or spatial sense".[82] While the law "does not make everyone responsible for everyone else",[83] it should not "veer towards an asocial view of responsibility".[84] We grasp this intuitively by standing in the defendant's shoes and reflecting on whether reasonable care should have been taken. In this manner, questions of what we owe each other as human beings are constantly implicated in the morality at the heart of the tort of negligence.[85] This can be explained and justified using the norms of friendship.[86] Friendship contains two norms: legitimate expectations and intrinsic worth.[87] The former demands that we recognise the claims we have on our friends and the reciprocal claims they make on us.[88] The latter informs us that friendship, "in which each is loved as an end, attests to the intrinsic worth of each person".[89] Law is also concerned with these two norms. The rights and obligations arising from a legal relationship have the nature of norms similar to those in friendship (i.e., legitimate expectations and reciprocity).[90] Once law recognises these norms in certain relations, they cannot be "denied in other relationships involving similar persons".[91] To illustrate, once the law holds that a duty of care exists between a doctor and a patient, it creates a set of rights and obligations between them. This set of rights and obligations should also exist between other doctors and their

78 Mark Turner and Giles Fauconnier, 'Conceptual Integration in Counterfactuals'. in Jean-Pierre Koenig (ed), *Discourse and Cognition: Bridging the Gap* (University of Chicago Press 1998).

79 Ruth MJ Bryne, 'Counterfactual Thinking: From Logic to Morality' (2017) 26(4) Current Directions in Psychological Science 314, 318–20; Nicole Van Hoeck, Patrick D Watson and Aron K Barbey, 'Cognitive Neuroscience of Human Counterfactual Reasoning' (2015) 9 Frontiers in Human Neuroscience 1.

80 Michael S Moore, *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics* (OUP 2009) 371.

81 Gary Low, 'Emphatic Plea for the Empathic Judge' (2018) 30 SAcLJ 97 [15].

82 John Gardner, *From Personal Life to Private Law* (n 64) 47.

83 Tan Seow Hon, *Justice as Friendship* (Ashgate 2015) 155.

84 ibid. See also Manderson (n 44) 36.

85 ibid 156; John CP Goldberg and Benjamin C Zipursky, 'The Restatement (Third) and the Place of Duty in Negligence Law' (2001) 54(3) Vanderbilt L Rev 657, 735; Samuel Scheffler, *Human Morality* (OUP 1992) 68–69.

86 Scheffler (n 85) 75–109.

87 Tan (n 83) 89; see *Toh Siew Kee* (n 7) [22].

88 ibid.

89 ibid.

90 ibid.

91 ibid.

patients. Moreover, law is also concerned with the "dignity of human beings in general".[92] In serving as a guide to human conduct, it is based on the conception of man as a "responsible agent with dignity".[93] Therefore, the norms in friendship can serve to justify law.[94] In doing so, it reflects the "relational nature of justice attach[ing] to"[95] the particular relationship between the parties.

(ii) Accommodating the Instrumental View within the Relational View

That said, the instrumental and relational views are not necessarily mutually exclusive.[96] The instrumental view can also be accommodated within a framework that is based on a relational view.[97] In saying that instrumentalist concerns can influence the relationship between the tortfeasor and claimant when social welfare considerations are considered in deciding whether a duty of care should be imposed,[98] I depart from Weinrib's view that arguments seeking to "have the law achieve goals external to the parties' relationship – whether instrumental, distributive, or economic – are all structurally inconsistent with fair and coherent determinations of liability".[99] As Gardner points out, legal recognition of this relationship between the parties is a question of distributive justice.[100] How, then, can this be consistent with a relational view of tort law which deals with interpersonal justice? The answer is apparent if one considers that policy reasons can modify the legitimate expectations of parties,[101] thereby affecting the bilateral relationship between such that this relationship cannot justifiably be recognised at law. Instrumentalist concerns of distributive justice are typically reflected in policy reasons which deal with collective welfare and social goals. The availability of insurance, which encapsulates the instrumentalist concern of loss-spreading, is one example.[102] To be clear, policy reasons feature in modifying the legitimate

92 ibid.

93 Lon Fuller, *The Morality of Law* (Yale University Press 1964).

94 Tan (n 83) 89.

95 ibid 92.

96 Marco Jimenez, 'Finding the Good in Holmes's Bad Man' (2011) 79 Fordham L Rev 2069, 2117–18.

97 See John Oberdiek, 'Method and Morality in the New Private Law of Torts' (2012) 125 Harvard L Rev Forum 189, 190–91; John Gardner, 'What is Tort Law For? Part 2. The Place of Distributive Justice' in John Oberdiek (ed), *Philosophical Foundations of the Law of Torts* (OUP 2018) 346.

98 *Spartan Steel & Alloys Ltd v Martin & Co (Contractors) Ltd* [1973] QB 27, 38; Tan (n 83) 89; John Gardner, 'What is Tort Law For? Part 2. The Place of Distributive Justice' (n 97); Andrew Robertson, 'On the Function of the Law of Negligence' (2013) 33(1) OJLS 31, 36–37.

99 Ernest Weinrib, 'Private Law and Public Right' (2011) 61 U of Toronto LJ 191, 192.

100 John Gardner. 'What is Tort Law For? Part 2. The Place of Distributive Justice' (n 97) 341.

101 ibid 159.

102 *Tan Juay Pah v Kimly Construction Pte Ltd and others* [2012] 2 SLR 549 [87]; *NTUC Foodfare* (n 15) [55]–[56].

expectations of parties.[103] The issue is not whether imposing a duty of care would result in increasing insurance premiums; rather, if insurance is available, both parties cannot legitimately expect that they can have recourse to tort as there is an insurance policy in play. And because they cannot legitimately expect to have recourse to tort, this justifies the court's non-recognition of the bilateral relationship at law.

How then does the relational view advanced above gel with the *Spandeck* framework? At the first stage, the concept of proximity establishes the bilateral relationship between tortfeasor and claimant. At the second stage, policy factors either favour or militate against the recognition of this bilateral relationship at law by modifying the legitimate expectations as between tortfeasor and claimant.

To be clear, the sort of policy reasoning at the second stage of *Spandeck* differs from that which Weinrib staunchly criticises.[104] Rather, it resembles the second notion of policy which Weinrib argues is not only "consistent with but also required by the general conception of duty".[105] In considering whether policy factors justify imposing a duty of care, the Singapore courts not only "explicate the legal meaning of that relationship in its particular circumstances"[106] but also demonstrate how it *modifies* the legitimate expectations parties have and, in so doing, provide a justification for imposing a duty of care.

Two implications follow from adopting a relational view. First, as alluded to, it illustrates the moral element within the tort of negligence: the breach of a duty is a wrong in and of itself. Second, and following from the first point, because the breach of a duty is a wrong, duties of care carry normative import. I deal with both points *seriatim*.

(iii) The Morality of a Duty of Care

The concept of a duty of care represents the moral element within the tort of negligence.[107] And, as explained earlier,[108] because distributive and policy criteria equally affect what both parties can legitimately expect or demand of each other, it also influences the moral relationship between them. Breach of this duty means that the tortfeasor has violated the moral relationship founded on equality

[103] ibid.
[104] Ernest Weinrib, 'The Disintegration of Duty' (n 10) 238–39.
[105] ibid 253.
[106] ibid.
[107] David Ibbetson, *A Historical Introduction to the Law of Obligations* (OUP 2001) 196; Avihay Dorfman, 'Can Tort Law be Moral?' 23(2) Ratio Juris 205, 210–11.
[108] See Section C (ii) "Accommodating the Instrumental View within the Relational View" above.

between both parties by risking the claimant's valuable interest.[109] Examples of these interests, which according to Perry are deserving of protection because they are central to human well-being, include, *inter alia*, life, health, dignity and "certain kinds of property interest".[110] This moral relationship recognises the rights people have "against interference [with their interests] by other persons".[111] Gardner labels this as "a raw moral duty" [112] that is distinguishable from a moral norm of corrective justice. Breach of this moral duty "creates a secondary duty to the same rightsholder".[113] Performance of this secondary duty reduces the "deficit in one's reason conformity that was left by one's non-performance"[114] of the original raw moral duty. Mapping this to the tort of negligence, breach of a duty of care is a breach of a moral duty owed to the claimant. The defendant has risked the plaintiff's valuable interest.[115] This therefore creates a *secondary* duty to repair the "deficit in one's conformity" with the duty owed. This secondary duty contains the moral norm of corrective justice; we are *obligated* to repair the wrong occasioned by the breach of our duty.[116] There are therefore two moral obligations: the original obligation that was breached and the secondary obligation to attempt to repair the breach of the original obligation.[117]

However, this "raw moral duty" has been obfuscated by instrumentalist views which focus on compensation for damage or loss as the only means of discharging this secondary obligation.[118] After all, actionable damage,[119] and

[109] John Oberdiek, 'The Moral Significance of Risking' (2012) 12 Legal Theory 339. See also Special Morality of Tort Law (n 66) 409. Deciding what interests are valuable and worthy of protection implicates our views on the good life. See Nicholas J McBride, "Tort Law and Human Flourishing" in Pitel, Neyers and Chamberlain (eds), *Tort Law: Challenging Orthodoxy* (Hart Publishing 2013) 34–57; J.M. Finnis, *Natural Law and Natural Rights* (OUP 2011) 81–97. McBride, however, disagrees with Finnis's conception of human flourishing.

[110] Stephen Perry, *Torts, Rights and Risk* (n 61) 54–55. Perry derives his list of interests from the idea that harm is a "relatively specific moral concept which requires that a person have suffered serious interference with one or more interests that are particularly important to human well-being".

[111] Stephen Perry, 'On the Relationship Between Corrective and Distributive Justice'. in Jeremy Horder (ed), *Oxford Essays in Jurisprudence, Fourth Series* (OUP 2002) 239.

[112] John Gardner, 'What is Tort Law For? Part 2. The Place of Distributive Justice' (n 97) 339.

[113] ibid 338; What is Tort Law for? Part 1 (n 25).

[114] ibid 339; What is Tort Law for? Part 1 (n 25) 34.

[115] John Oberdiek, 'The Moral Significance of Risking' (n 108).

[116] John Gardner, 'Backwards and Forwards with Tort Law' 29 – 30 < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1397107> accessed 3 May 2021.

[117] John Gardner, 'Torts and Other Wrongs' University of Oxford Legal Research Paper Series (August 2011) 25.

[118] Jules Coleman, 'Tort Law and Tort Theory, Preliminary Reflections on Method'. in Gerald Postema (ed), *Philosophy and the Law of Torts* (Cambridge University Press 2002) 189.

[119] Donal Nolan, 'New Forms of Damage in Negligence' (2007) 70 MLR 59. See Christian Witting, *Street on Torts* (OUP 2015) 5–7.

causation must also be proven.[120] This reflects the divergence between a moral wrong arising from the breach of a duty owed *simpliciter* and a legal wrong. The focus on legal wrongs glosses over the consequences of breaching a duty of care, which is itself a wrongful act, and our obligation to set things right, regardless of whether harm has been occasioned. As Radzik posits, a wrongful act may not necessarily result in harm because the harm has either been (a) repaired by a third party or (b) avoided through sheer luck.[121] However, absence of harm occasioned does not detract from the fact that the act, or failure to act, itself is *wrong* or that we no longer have a moral obligation to remedy our breach of the duty of care. Availability of a legal remedy does not necessarily absolve us of the secondary obligation of repair. Money cannot fix everything, including repairing moral wrongs.[122] An apology, however, might suffice.[123]

This divergence is evident from cases where courts dismissed the claim on grounds that causation was not proven, despite finding a breach of the duty owed. In *Gregg v Scott*,[124] Lord Nicholls recognised that it was irrational to hold that a patient could only claim damages arising from a loss of chance when he had "lost a 55% [chance] of recovery but not a 45% [chance] of recovery".[125] In *both* cases, the doctor was "in breach of his duty to the patient".[126] Disallowing the claim on the difference between a 45% chance of recovery and a 55% chance would result in an "a duty [devoid] of content".[127] Clearly Lord Nicholls recognised that a doctor's breach of the duty was a *wrong*.[128] To his mind, a wrong occasioned should entitle the claimant to a remedy, or otherwise would strip the duty of care of any meaning. In so doing, Lord Nicholls seemed to equate legal wrongs with moral

[120] Michael Jones (ed), *Clerk and Lindsell on Torts* (n 75) [2-01].

[121] Linda Radzik, 'Tort Processes and Relational Repair' in John Oberdiek (ed), *Philosophical Foundations of the Law of Torts* (OUP 2018) 245–48. See also CP Goldberg and Benjamin C Zipursky, 'Torts as Wrongs' (2010) 88(5) Texas L Rev 917, 935.

[122] Compensation may sometimes suffice to repair a moral wrong, see William Lucy, *Philosophy of Private Law* (OUP 2007) 314–16.

[123] Linda Radzik, 'Tort Processes and Relational Repair' in John Oberdiek (ed), *Philosophical Foundations of the Law of Torts* (OUP 2018) 238.

[124] [2005] 2 AC 176.

[125] ibid [3].

[126] ibid.

[127] ibid [4].

[128] John Oberdiek, 'The Moral Significance of Risking' (n 109) 350–54.

wrongs when both are distinct. This may explain why the moral aspect of a duty of care has been overlooked.

(iv) The Normative Dimension of Duties Owed

Second, because the breach of a duty of care is, in and of itself, a wrong, duties of care carry normative import. This can be gleaned from the "critical reflective attitude"[129] of society. Adopting the internal observer's viewpoint,[130] this internal aspect is manifest in deviation from the rule.[131] This is expressed in the language of normative vocabulary. For instance, one might say: "A *ought* to have taken reasonable care in this situation" or that "A was *wrong* for failing to take reasonable care in such a situation". Because the breach of this duty of care is a wrong, it carries normative import from an internal viewpoint. As McBride put it, "if A is said to owe B a duty to take care not to do x in a given situation, A will *actually* have a duty to take care not to do x, which duty will have been imposed on A for B's benefit".[132] It is in this manner that the law serves as a guide to human conduct[133] and, therefore, judicial pronouncements of the existence of a duty of care hold normative force. If a court finds that a duty of care is owed in a particular situation, and a breach of a duty of care is a wrong, then other people *ought* to take reasonable care in similar circumstances. This reflects the relational view of tort explained above; there is an expectation that others would also take reasonable care in similar circumstances as well. In the process, rules influencing the critical reflective attitude of members of society are laid down, signalling that a duty of care is owed in that particular situation.[134]

That this view reflects how participants practise and understand tort law is apparent from the judgements.[135] In *Noor Azlin Binte Abdul Rahman v Changi General Hospital Pte Ltd* ('Noor Azlin'),[136] the court held that the senior respiratory physician

[129] Scott J Shapiro, 'What is the Internal Point of View' (2006) 75 Fordham L Rev 1157, 1164–65.

[130] HLA Hart, *The Concept of Law* (3rd edn, OUP 2012) 88–90.

[131] See also Philip Pettit, 'Social Norms and the Internal Point of View: An Elaboration of Hart's Genealogy of Law' (2019) 39(2) OJLS 229, 245, 251.

[132] Nicholas J McBride, 'Duties of Care – Do they Really Exist?' (2004) 24(3) OJLS 417 (emphasis added).

[133] John CP Goldberg, Benjamin C Zipursky, 'Seeing Tort Law from the Internal Point of View: Holmes and Hart on Legal Duties' (2006) 75(3) Fordham L Rev 1563, 1577; Christian Witting, 'Duty of Care: An Analytical Approach' (2005) 25(1) OJLS 33, 42.

[134] Stephen A Smith, 'The Normativity of Private Law' (2011) 31(2) OJLS 215, 231.

[135] John CP Goldberg, Benjamin C Zipursky, 'Seeing Tort Law from the Internal Point of View: Holmes and Hart on Legal Duties' (2006) 75(3) Fordham L Rev 1563, 1575; John CP Goldberg, Benjamin C Zipursky, 'The Restatement (Third) and the Place of Duty in Negligence Law' (2001) 54(3) Vanderbilt L Rev 657, 732.

[136] [2019] 1 SLR 834.

that examined the plaintiff had breached his duty of care.[137] However, causation was not established on the facts. In *Yeo Peng Hock v Pai Lily* ('*Yeo Peng Hock*'),[138] the court similarly held that the doctor had breached his duty of care in failing to send the patient to the Accident & Emergency department. However, the claim failed as causation was not established. The language used expresses the normative dimension of the duty of care. In *Noor Azlin*, the court opined that the senior respiratory physician "*ought* to have taken the more cautious route of scheduling a follow-up" if he was unsure of the diagnosis.[139] Similarly, in *Yeo Peng Hock*, the court concurred with the trial judge's finding that "any competent GP *would* have advised his patient to go immediately to a hospital".[140] This demonstrates McBride's point: If A is said to owe B a duty, A *actually* has a duty to take reasonable care in relation to B. The language of the judgement reflects that this duty exists, illuminating its normative dimension in the form that the defendant 'ought' to have done X or that any reasonable man in that position 'would' have done X. Although causation in both cases was not established, the finding of a breach of a duty of care demonstrates that a duty of care is indeed owed under such circumstances and reflects the court's opinion as to what must be done to discharge that standard of care. Consequently, the finding of a duty of care clearly has a normative dimension.

Because a duty of care carries normative import, it is unsurprising that judges have relied on it *qua* control mechanism.[141] Properly understood, the elements of the tort of negligence may overlap,[142] but should remain distinct inquiries. However, in utilising the duty of care as a control mechanism, the court has collapsed the analysis. In the UK, for example, judges have preferred to treat cases involving a breach of the standard of care as cases where no duty of care exists.[143] The case of *Darnley v Croydon Health Services* illustrates this.[144] The claimant was struck on the head after unknown assailants attacked him. He went to the hospital. He was informed by the receptionist that the waiting time was approximately 4–5 hours and was told to wait. After waiting for 20 minutes, he went home. His condition deteriorated. He was sent back to the hospital by ambulance. Unfortunately, by then, he had already suffered serious and permanent injury because of the delay in

[137] ibid [91].
[138] [2001] 3 SLR(R) 555.
[139] [2019] 1 SLR 834 [89] (emphasis added).
[140] [2001] 3 SLR(R) 555 [18] (emphasis added).
[141] Ken Oliphant, 'Against Certainty in Tort Law'. in Stephen GA Pitel, Jason W Neyers and Erika Chamberlain (eds), *Tort Law: Challenging Orthodoxy* (Hart Publishing 2013) 5.
[142] James Goudkamp, 'Breach of Duty: A Disappearing Element of the Action in Negligence?' (2017) Cambridge LJ 480, 480.
[143] ibid 481.
[144] [2018] QB 783.

treatment. The UK Court of Appeal ('UKCA'), instead of focussing on the breach of duty owed, focussed on whether a duty was even owed in the first place. This was surprising as "*Darnley* was completely lacking in features that could possibly be thought to have given rise to any duty issue".[145]

While *Darnley* was overturned on appeal,[146] the UKCA's decision remains highly unsatisfactory as it distorts the duty of care analysis by examining whether a factual duty, which deals with whether harm to the plaintiff was a reasonably foreseeable consequence of the defendant's conduct,[147] exists. However, because remoteness already deals with the same question, defining a duty of care in this manner renders it otiose.[148] To avoid this, the duty of care should be concerned with notional duties. The question is a normative one: should the law of negligence be applied to the present case?[149] To answer that question, the courts have relied on proximity *qua* concept.

(v) A Desire for Certainty

In summary, much of the confusion surrounding the proximity requirement can be traced to the linguistic usage of the word 'proximity'. Utilising Zipursky's pragmatic conceptualism, "proximity" and the *Spandeck* framework denote a relational view of tort law which also encompasses typically instrumentalist concerns. One might further note that the overriding concern with compensation has obfuscated the relational view of tort and the significance of a duty of care; *viz.* that the breach of a duty of care is a wrong. The conflation between notional and factual duties is problematic. Attempting to rein in liability, courts conflate the duty of care inquiry with other elements of negligence. One might attribute this to the desire for certainty over the outcomes of individual cases.[150] Given the normative dimensions of finding that there is a duty of care, courts are naturally wary of sending a wrong signal to society. This attitude can be traced to the tentative nature in which the tort of negligence was developed.[151]

That said, the lack of certainty in terms of *outcomes* is certainly not deleterious. Courts should "state the principles according to which a duty of care should be

[145] James Goudkamp, 'Breach of Duty: A Disappearing Element of the Action in Negligence?' (n 140) 482.

[146] *Darnley v Croydon Health Services NHS Trust* [2018] UKSC 50.

[147] Plunkett (n 26) 82; Clerk and Lindsell on Tort (n 75) [8-07]; Colin Liew, 'Keeping it Spick and Spandeck: A Singaporean Approach' (2012) 20 Torts LJ, 9–10.

[148] Plunkett (n 26) 89.

[149] ibid 111; Michael Jones (ed), *Clerk and Lindsell on Torts* (n 75) [8-06].

[150] Ken Oliphant, 'Against Certainty in Tort Law' (n 141) 4.

[151] *Langridge v Levy* (1842) 150 ER 863; *Winterbottom v Wright* (1842) 10 M & W 109; *Longmeid v Holliday* (1851) 6 Ex 761; *George v Skivington* (1869) LR 5 Ex Rep 1; *Heaven v Pender* (1883) 11 QBD 503.

determined" and "engage in a flexible weighing up of all normatively relevant factors".[152] As I have sought to demonstrate, the key inquiry where the duty of care is concerned is the concept of proximity.[153] This is because it accurately reflects the underlying conceptual understanding that the tort of negligence is primarily relational.

We turn now to examine how *Spandeck* has been applied in practice, focussing mainly on Court of Appeal judgements because of the authoritativeness of its rulings, to determine if it indeed reflects the understanding of proximity that reflects the relational view as sketched out above. Where *Spandeck* departs from the relational view, I propose changes.

IV. *Spandeck*'s Conceptual Coherence

A. Factual foreseeability

(i) Case Law

In this Part, I examine whether factual foreseeability is consistent with a relational view and its logical coherence with the other elements of the *Spandeck* framework.[154] *Spandeck* conceptualised factual foreseeability as a threshold test. If the facts did not evince that it was foreseeable that the plaintiff would suffer harm if the defendant failed to take reasonable care, this threshold requirement would not be crossed.[155] *Ngiam Kong Seng v Lim Chiew Hock* ('*Ngiam*')[156] is one example. The first appellant was involved in a traffic accident allegedly caused by the respondent who represented himself as a good Samaritan that rendered aid to the first appellant. Consequently, the second appellant developed feelings of gratitude towards him[157] but, upon discovering the respondent's role in the accident, she developed depression and suicidal tendencies resulting from a sense of betrayal.[158] In considering whether the respondent owed a duty of care to the second appellant, the court held that the factual foreseeability requirement was not satisfied as "it was not reasonably foreseeable that the mere communication of the information in question without more could result in harm to a party".[159] Nevertheless, the court proceeded to analyse the existence of a duty of care based on the first stage of the

[152] Ken Oliphant, 'Against Certainty in Tort Law' (n 141) 18.
[153] See Christian Witting, 'Duty of Care: An Analytical Approach' (2005) 25(1) OJLS 33, 38–42.
[154] See Gary Chan, *The Law of Torts in Singapore* (2nd edn, Academy Publishing 2016) [03.041].
[155] *Spandeck* (n 7) [75]–[76].
[156] [2008] 3 SLR(R) 674.
[157] ibid [7].
[158] ibid.
[159] ibid [132].

Spandeck framework. Absent a professional relationship between the plaintiff and the defendant (as was the case in *Ngiam*), there was no duty of care not to pass on information that could cause psychiatric shock.[160]

(ii) Problems

This conceptualisation of factual foreseeability is problematic. Distinguishing between the foreseeability of harm and the foreseeability of the *type* of harm is hardly possible.[161] In pointing out that harm to the second appellant was unforeseeable, the court in *Ngiam* discussed the type of harm, *viz.*, psychiatric harm. This confuses the duty of care inquiry with the remoteness rule, despite the warning in *Spandeck*.[162] The case of *AYW v AYX* ('*AYW*')[163] demonstrates this. In *AYW*, the High Court struck out the claim on the ground that it did not meet the threshold requirement of factual foreseeability.[164] The plaintiff in *AYW* sued the school in negligence for failing to deal with alleged acts of bullying. Considering whether the school owed a duty of care to the plaintiff, the court held that, while schools owed a duty of care towards their pupils, they had no duty to take reasonable care in protecting students from *all* types of harm.[165] The duty of care did not extend to intervening in the "bullying" as alleged in the statement of claim. In totality, the court opined that it was not factually foreseeable that the plaintiff would suffer any physical/psychiatric injury or economic loss arising from the bullying. There was "no suggestion of a persistent pattern of physical gestures (let alone threatening gestures) over a period of time [that would] give rise to a foreseeable risk of harm if steps were not taken to intervene".[166] Moreover, the court in *AYW* also seemed to equate the failure to cross the factual foreseeability threshold with grounds for striking out.[167]

This conceptualisation of factual foreseeability puts the cart before the horse. A duty of care can exist despite the damage being too remote. This understanding of factual foreseeability collapses the duty of care inquiry into a single stage: was the damage caused a reasonably foreseeable consequence of the defendant's actions? [168] Because the court also held that schools owed a duty to take

[160] ibid [142].
[161] Plunkett (n 26) 98–104.
[162] *Spandeck* (n 7) [89].
[163] [2016] 1 SLR 1183.
[164] ibid [91], [94].
[165] ibid [69]–[70]. The court was not referring to non-delegable duties. It considered this under the heading: "Did the School owe the Plaintiff a duty of care".
[166] ibid [86].
[167] ibid [91], [94].
[168] Plunkett (n 26) 84–89.

reasonable care to protect students, the real issue in *AYW* was remoteness rather than the existence of a duty of care.

(iii) Clarifying Factual Foreseeability

Therefore, at the factual foreseeability stage, the court examines the *facts* to determine if it was foreseeable to the defendant that the plaintiff's *interest* would be endangered if reasonable care were not taken.[169] This is consistent with a relational view. If it were foreseeable that the defendant's actions would endanger the interests of a class of people to which the plaintiff belongs ('foreseeability requirement'),[170] this would create a legitimate expectation that he takes reasonable care when acting. Minimally, the foreseeability to the plaintiff that his actions might affect the interests of a class of people to which the plaintiff belongs is the ingredient needed to indicate that a potential legal relationship exists between both the plaintiff and defendant.

As Plunkett argues, citing *Smith* as an example,[171] requiring foreseeability that the plaintiff's interest might be endangered does not encounter the same problems as requiring foreseeability of harm to the plaintiff. In that case, the defendant train company had allowed dry grass to accumulate near its railway tracks.[172] Sparks from a passing locomotive ignited the grass. The fire spread. The adjoining stubble field and the plaintiff's cottage were destroyed. According to the interest theory, because the cottage was located quite a distance from the tracks, and the plaintiff did not own the stubble field, the defendant "had not been negligent vis-à-vis the plaintiff's property interest in his cottage".[173] Plunkett argues that difficulties arise if we hypothesise that the plaintiff had also owned the stubble field as he would be able to claim for damage to the cottage as his property interest was affected. This would be a "capricious result" as the plaintiff's claim depended on who owned the stubble field.[174] One might attempt to distinguish an interest in the stubble field as being different from the interest in the cottage, but this requires flexibility and discretion. There is therefore no meaningful distinction between "interest and kinds of harm".[175]

However, applying the foreseeability requirement based on the interest theory, the plaintiff might not be able to claim for the damage to the cottage even if

[169] John Gardner, *From Personal Life to Private Law* (n 63) 50–51.
[170] Gary Chan, *Law of Torts in Singapore* (n 154) [03.042]–[03.043]. See Allan Beever, *Rediscovering the Law of Negligence* (Hart Publishing 2007) 133.
[171] Plunkett (n 26) 102.
[172] ibid.
[173] ibid.
[174] ibid.
[175] ibid 103.

he owned the stubble field. The defendant would owe a duty of care as it would be foreseeable that the plaintiff's property interest, which covers both the stubble field and the cottage, would be affected if they failed to take reasonable care. However, the claim for damage to the cottage can be denied on grounds of remoteness. One might argue that it was unforeseeable that the fire would spread that far and damage the cottage. Thus conceived, our foreseeability requirement at the duty of care stage examines whether the plaintiff's interest has been endangered. Because factual foreseeability is a threshold requirement, the plaintiff's interest should be broadly construed and dealt with at a high level of generality. The *extent* to which the plaintiff's interest has been injured is reflected by the remoteness inquiry which deals with the foreseeability of harm. Here, damage to the stubble field was foreseeable. Damage to the cottage was not. Therefore, the plaintiff's property interest (in both the stubble field and cottage) was not *wholly* damaged.

The role of factual foreseeability, then, is simply this: if the facts do not evince that it was foreseeable that the plaintiff's interest would be endangered, there is no need to apply the *Spandeck* framework. That said, it is good practice to proceed with the proximity analysis under the first stage of *Spandeck* as it provides valuable guidance as to when the factual foreseeability threshold is crossed, and when a duty of care is established.[176] So conceptualised, factual foreseeability weeds out cases where there is no relationship between the parties at all and the law of negligence simply does not apply. Factual foreseeability can therefore serve as grounds for striking out. If the facts do not even disclose the existence of a relationship between the parties, which is the crux of negligence, it is plain and obvious that the claim has no substance.[177] Applying the reformulated conception of factual foreseeability to *AYW*, the claim would not have been struck out on the ground that the factual foreseeability threshold was not met. Arguably, it was foreseeable on the facts that the plaintiff's interest in her well-being or dignity would have been put at risk by the defendant's failure to take reasonable care in stopping the alleged acts of bullying. The claim, however, could have been struck out on grounds of *remoteness* instead.[178]

B. Stage 1: legal proximity

While cases have alluded to the concept of proximity having some substantive content,[179] little has been said about what this substantive content is. Earlier, we explained how proximity reflected a relational view of tort law based on corrective

[176] Ngiam (n 156) [32]. The court proceeded on the assumption that factual foreseeability could be established.

[177] *Gabriel Peter & Partners (suing as a firm) v Wee Chong Jin and others* [1997] 3 SLR(R) 649 [21]–[22].

[178] *AYW v AYX* [2016] 1 SLR 1183 [89].

[179] *Spandeck* (n 7) [80].

justice. Having established that factual foreseeability is a filtering mechanism, the analysis at the legal proximity stage can be conceptualised accordingly: the court should explain *why* it is foreseeable that the defendant's actions would endanger the plaintiff's interests. This gives the duty of care its normative dimension by justifying why the plaintiff had to take reasonable care. We might express it as follows: "if it is foreseeable that the defendant's actions would have endangered the plaintiff's interest, then he *ought* to have taken more care in acting".

Notably, *Spandeck* highlighted that this stage was to be applied incrementally. This incremental approach has been described as a disguise for policy reasoning.[180] However, properly understood, the incremental approach is nothing more than applying the common law method of analogical reasoning. Cases from other common law jurisdictions can be used if the facts are indicative of the ways in which the plaintiff's interest may be endangered by the defendant.[181] Based on the manner in which *Spandeck* has been applied, the ways in which the plaintiff's interest may be endangered by the defendant have been categorised under the following proximity factors of VAR-R,[182] *Sutherland* proximities,[183] and knowledge.[184]

Usage of proximity factors also reflects the relational view described above. Take, for instance, VAR-R, which was defined in *Go Dante Yap v Bank Austria Creditanstalt AG*.[185] The plaintiff in that case had some investments with the defendant bank that went south. He sued, alleging that the bank owed him a duty of care in relation to the provision of services and executing his instructions.[186] The court held that, notwithstanding the contractual framework, there was VAR-R that sufficed to establish sufficient proximity between the parties. This was because the defendant bank had "accepted the [plaintiff] as someone whose money and assets were under its control and on whose behalf it could and was expected to expend considerable sums to acquire various investments".[187] By "offering private banking and wealth-management facilities", the bank "held itself out as possessing special skill or expertise".[188] Relying on this skill and expertise, the plaintiff allowed the bank to act on his behalf. Reliance on the defendant's skill, coupled with the defendant's acceptance of that reliance by assuming responsibility, means that the actions of the defendant would directly impact the plaintiff's valuable interest.

[180] Beever (n 170) 183–89.
[181] David Tan and Goh Yihan, 'The Promise of Universality' (n 8) 18.
[182] *NTUC Foodfare* (n 15) [40]; *Animal Concerns* (n 13) [60].
[183] *Spandeck* (n 7) [78]–[79].
[184] See David Tan and Goh Yihan, 'The Promise of Universality' (n 8) 26.
[185] [2011] 4 SLR 559; Justin Tan, 'Proximity as Reasonable Expectations' (n 36) 151.
[186] ibid [2].
[187] ibid [35].
[188] ibid.

There is a legitimate expectation that the defendant, being in a position where his actions could affect the plaintiff's valuable interest, would act with reasonable care to avoid endangering it.

We turn next to the *Sutherland* proximities. Causal proximity refers to the "causal connection" between the defendant's actions and the harm suffered by the plaintiff.[189] However, this is different from the idea of a causal connection between the defendant's actions and the risk posed to the plaintiff's valuable interest which goes towards establishing the breach of a duty. As explained above, the breach of a duty is a moral wrong that is distinct from a legal wrong. There must therefore be a causal link between the defendant's actions and the risk posed to the plaintiff's valuable interest. Here, we are concerned with explaining *why* the defendant's actions could endanger the plaintiff's interest; a causal link between the defendant's actions and harm suffered by the plaintiff clearly indicates that the defendant's actions could endanger the plaintiff's interest.

Causal proximity, however, is not the only way of explicating this. Take, for example, physical proximity, as discussed in *Animal Concerns Research & Education Society v Tan Boon Kwee*,[190] which refers to the closeness in time and space between the plaintiff and defendant. The plaintiff hired the defendant to serve as the site supervisor in the construction of an animal shelter. The shelter was not constructed according to specified building plans. Wood chips used to level the site decomposed, necessitating remedial action on the plaintiff's part.[191] The plaintiff sued, alleging that the defendant had "failed to supervise the levelling of the site" and that the "wood chips were [un]suitable landfill material".[192] The court held that there was physical proximity between the parties because the defendant was physically present at the site.[193] This reflects the relational view. Because the defendant was physically present, he could act to mitigate or eliminate the risk posed to the plaintiff's interest.

The last of the *Sutherland* proximities, circumstantial proximity, refers to the parties' "factual relationship".[194] In *See Toh Siew Kee v Ho Ah Lam Ferrocement (Pte) Ltd*,[195] VK Rajah JA held that circumstantial proximity is "tautologically present in the occupier [and] lawful entrant relationship".[196] Clearly, the occupier's failure to maintain his property could undoubtedly risk the lawful entrant's interest in

[189] Justin Tan, 'Proximity as Reasonable Expectations' (n 36) 149.
[190] [2011] 2 SLR 146 (n 13); see also *See* Toh Siew Kee (n 7).
[191] ibid [7].
[192] ibid [8].
[193] ibid [37].
[194] Justin Tan, 'Proximity as Reasonable Expectations' (n 36) 149.
[195] [2013] 3 SLR 284.
[196] ibid [80].

bodily integrity. Similarly, in *Ramesh s/o Krishnan v AXA Life Insurance Singapore Pte Ltd*,[197] the High Court held that circumstantial proximity was established because of the past employer-employee relationship between the plaintiff and defendant.[198] This reflects the relational view because, by being in an employer-employee or occupier-lawful entrant relationship, the parties are placed in a position whereby their actions could affect each other's interests.

Finally, we turn to knowledge, which was used as a proximity factor in *Anwar Patrick Adrian and another v Ng Chong & Hue LLC* ('*Anwar*').[199] In *Anwar*, the defendant solicitor was hired by the plaintiff's father to restructure debts owed to the bank. The father told the solicitor that he did not want his sons to be personally liable for the debts.[200] However, the defendant solicitor failed to point out the presence of a clause in the Security Documents, under which the plaintiffs had agreed to personally guarantee their father's debts.[201] The Bank claimed against the plaintiffs under this clause. The plaintiffs subsequently sued the defendant solicitor for failing to inform them that such a clause was present. The court held that the defendant's knowledge of affairs "support[ed] a finding of proximity".[202] The defendant *knew* that he was being retained to ensure that the plaintiff's interests were protected.[203] There is therefore a legitimate expectation that the defendant solicitor would act with reasonable care, lest his negligence endanger the plaintiff's financial interest. Simply put, if we *know* that our actions could potentially place another's interest at risk, then the onus is on us to act with reasonable care. Undoubtedly, applying the Golden Rule, we would expect the same of others.

It is, however, important to note that the usage of proximity factors differs from the 'pockets approach'. Under that approach, cases are not "decided according to broad general tests or principles which underlie all duty cases".[204] Instead, reference is made to the underlying reasons for the outcome of cases with similar factual matrices.[205] However, the Singapore courts have applied more than one proximity factor in cases.[206] Moreover, as we have sought to argue, the core of *Spandeck* lies in proximity and the relational view. In that light, these proximity factors represent more of a categorical approach to the question of a notional

[197] [2015] 4 SLR 1.
[198] ibid [243].
[199] [2014] 3 SLR 761.
[200] ibid [15].
[201] ibid [25].
[202] ibid [148].
[203] ibid.
[204] Plunkett (n 26) 70.
[205] ibid.
[206] *NTUC Foodfare* (n 15) [47], [48], [50]; *Ramesh* (n 57) [251]–[255].

duty.[207] This method of analysis, coupled with the use of precedent, [208] allows judges to justify their finding on a duty of care. After all, both parties come to court, believing that they *have* a legitimate claim (even more so if the claim is not struck out at the interlocutory stage).[209] Justice must not only be done, but must also be *seen* to be done by explaining,[210] in clear and principled terms using common law reasoning, the conclusion reached at the duty of care stage.

This allows us to see the three components of the *Spandeck* framework as separate, yet logically linked stages. Factual foreseeability deals with the sufficiency of the facts to facilitate the duty of care inquiry. It has no normative force, unlike Stage I of *Spandeck*. One cannot derive an 'ought' from an 'is';[211] a "mere appeal to the facts alone" cannot justify the imposition of a duty of care.[212] Indeed, duties of care are imposed by law.[213] Applying common law reasoning at Stage I of *Spandeck*, we would, with reference to previous cases, infer that a duty of care should be imposed where material facts A and B are present.[214] Given the similarity of the present facts to material facts A and B, we can conclude that a duty of care *should* be imposed in this situation.[215] Therefore, duties of care do have a "normative dimension".[216] This further allows us to distinguish the proximity analysis from the assessment of policy considerations whilst recognising how both stages can interact.[217] As explained above, policy considerations can modify parties' legitimate expectations. Stage I of *Spandeck* sketches out what these legitimate expectations should be, with reference to previous cases. At the policy stage, one considers if these legitimate expectations have been modified such that the law of negligence should not recognise the bilateral relationship between tortfeasor and claimant.

However, one clarification must be made in relation to the operation of indeterminate liability *qua* policy consideration.[218] In *NTUC Foodfare*, the

[207] Plunkett (n 26) 140.

[208] Keith Stanton, 'Decision-making in the tort of negligence in the House of Lords' (2007) 15 Tort L Rev 93, 94.

[209] Beever (n 170) 186–87.

[210] *R v Sussex Justices, Ex parte McCarthy* [1924] 1 KB 256.

[211] Rachel Cohon, 'Hume's Moral Philosophy' (Stanford Encyclopaedia of Philosophy, 20 August 2018) <https://plato.stanford.edu/entries/hume-moral/> accessed 14 November 2019.

[212] Andrew Phang, Cheng Lim Saw, and Gary Chan, 'Of Precedent, Theory and Practice - The Case for a Return to Anns' (n 28) 46.

[213] *Go Dante Yap* (n 58) [19].

[214] See J Montrose, 'The Ratio Decidendi of a Case' (1957) 20 MLR 587.

[215] ibid.

[216] Andrew Phang, Cheng Lim Saw, and Gary Chan, 'Of Precedent, Theory and Practice - The Case for a Return to Anns' (n 28) 45.

[217] ibid 54.

[218] Andrew Robertson, 'Policy-based reasoning in duty of care cases' (2013) 33(1) Legal Studies 119, 122.

court highlighted that the concept of proximity dealt with the question of indeterminate liability to an indeterminate class by "restrict[ing] recovery to a reasonably determinate class of persons".[219] However, indeterminate liability could feature under the policy stage.[220] While considerations of proximity and policy may overlap,[221] it seems illogical to consider the question of indeterminate liability to an indeterminate class as a policy factor. Logically, the proximity requirement eliminates this as a policy consideration. It would be illogical to hold that there is sufficient proximity between the parties and then to proceed to consider indeterminate liability to an indeterminate class under the policy rubric. Policy, then, necessarily deals with the question of indeterminate liability for an indeterminate amount. The inquiry here is slightly different from that in remoteness, which examines the foreseeability of the type of damage from the defendant's perspective. Policy situates the duty of care inquiry within the broader context of society: Could other people in a similar position to the defendant be said to have assumed the risk of indeterminate liability for an indeterminate amount in so acting? *Spandeck*, thus conceptualised, is logically coherent. Its three components are distinct and logically related to each other.

C. Stage 2: policy considerations

While policy is considered separately from proximity, we do not attempt to draw the same principle-policy divide as Lord Reed did in *Robinson*.[222] Lord Reed opined that policy reasons should only be applied to novel cases and not to cases falling within principles of the law of negligence as established through precedent. However, as argued above, policy considerations feature in the duty of care analysis by modifying the legitimate expectations of the parties. *Spandeck* recognised the role of policy factors in the duty of care analysis.[223] While it is hardly possible to wrest apart policy from principle,[224] separating the inquiry allows the court to be candid with policy reasoning to "avoid giving the impression that there [are] unexpressed

[219] *NTUC Foodfare* (n 15) [43].
[220] David Tan and Goh Yihan, 'The Promise of Universality' (n 8) [43].
[221] Andrew Phang, Cheng Lim Saw, and Gary Chan, 'Of Precedent, Theory and Practice - The Case for a Return to Anns' (n 28) 54.
[222] *Robinson* (n 2) [27].
[223] *Spandeck* (n 7) [84].
[224] Kenny Chng, Gary Chan and Goh Yihan, 'A Novel Development of Tort Law: Robinson v Chief Constable of West Yorkshire Police' (2019) 25 Torts LJ 184, 190–93.

motives [in] finding for or against a duty".[225] With this in mind, we explain how policy factors modify the legitimate expectations of parties.

One example of policy reasoning is the clash between a contractual duty and a tortious duty. In *Spandeck*, the policy reason for not imposing a duty of care was the need for caution before imposing a tortious duty onto a relationship which the parties had already chosen to regulate via contract.[226] This means that, in assessing the legitimate expectations between the parties, one should consider, in assessing whether the law of negligence *should* apply, that both parties had chosen to regulate their relationship via contract, having considered it more economically efficient to do so. In *Spandeck*, this was the case as the contract between the parties allowed the plaintiff claim to proceed under arbitration proceedings against the defendant. However, the court also concluded that there was no proximity for the very same reason: the presence of the arbitration clause.[227] The overlap between proximity and policy here is not problematic for two reasons. At a superficial level, it illustrates the need to be candid about policy considerations. Sans the policy stage, critics might argue that the court's finding of no duty in *Spandeck* was based on the policy ground that tortious duties should not be superimposed onto a contractual framework. Conceptually, the policy stage allows the court to articulate policy considerations inherent in the duty of care analysis and explain *why* a duty of care should not be imposed in the present case – the presence of a contract means that both parties should expect that their relations be governed by contract, rather than tort.

The consideration of statutory frameworks is another example. In *Jurong Primewide Pte Ltd v Moh Seng Cranes Pte Ltd*,[228] the court held that consideration of the "underlying statutory scheme and parliament['s] intention" is done at the policy stage of the *Spandeck* framework. The statutory framework must be considered because common law duties should not undermine the "effectiveness of duties imposed by the statute",[229] or "distort the focus of the statutory decision-making process" and "the performance of the functions of the statutory body".[230] Where a statute conflicts with a common law rule, the statute should prevail. Consideration of statutory frameworks modifies the parties' legitimate expectations because the defendant should have acted in accordance with the statutory framework. Similarly, the plaintiff will likely expect the same of the defendant, affecting the bilateral relationship such that the law cannot justifiably recognise a duty of care in this

[225] *Spandeck* (n 7) [85].
[226] ibid [101], [114].
[227] ibid [83].
[228] [2014] 2 SLR 360.
[229] Gary Chan, *Law of Torts in Singapore* (n 154) [05.082].
[230] ibid.

instance. After all, if the imposition of a duty of care is a problem of distributive justice, why should the tort of negligence apply when the statute already provides a solution? Therefore, in assessing the parties' legitimate expectations on this view, one should consider that the plaintiff can resort to the statute or to a claim for breach of a statutory duty as a remedy. Similarly, the defendant is likely to expect this of the plaintiff.

In summary, the threshold requirement of factual foreseeability is a filter mechanism; the court must examine the facts to determine whether it was foreseeable that the plaintiff's interests would be endangered. If this threshold requirement is met, the re-conceptualised *Spandeck* framework applies:

Stage I: Legal proximity requires the court to explain *why* the defendant's actions could have endangered the plaintiff's interests.

Stage II: Policy factors affecting the legitimate expectations of the parties are considered. This affects the overall analysis as to whether there should be a duty of care.

V. Conclusion

This article has sought to demonstrate, in two major parts, that the modified *Spandeck* framework is rooted in the concept of proximity which reflects the relational view at the heart of tort law. The first half of this article began with a brief description of the *Spandeck* framework before diving in to explain how its major components were consistent with the conceptual foundations of tort law. Key support to the argument is drawing the distinction between proximity *qua* descriptor and proximity *qua* concept. Once this distinction is grasped, it becomes clear that, insofar as Singaporean jurisprudence is concerned, usage of "proximity" refers to underlying tort law concepts, *viz.*, the relational view. This has important implications for tort law, namely that the duty of care is an important and distinct point of analysis in the tort of negligence, and that a duty is still owed, even though the plaintiff may not be able to demonstrate a breach of the standard of care and causation. At a more fundamental level, it reveals the moral implications of a duty of care, and what we owe to each other as human beings.

Building upon the analysis in the first half of this article, the second half assessed whether the *Spandeck* framework was consistent with the underlying conceptual foundation of tort law. Although largely consistent, tweaks need to be made to how the factual foreseeability stage is understood and applied. Having argued that general frameworks, such as *Spandeck*, can provide a principled analysis for assessing whether there is a duty of care, it is hoped that this will spark a reconsideration of such frameworks post-*Robinson*.