European Journal of International Law

Infecting the Mind: Establishing Responsibility for Transboundary Disinformation

Henning Lahmann*

Abstract

This article examines the legal issues concerning the establishment of responsibility for an internationally wrongful act in the context of transboundary disinformation. In light of the unprecedented surge of potentially dangerous health disinformation throughout the COVID-19 pandemic, there is growing consensus among academics and states that influence campaigns that utilize false or misleading information may qualify as a violation of international law, amounting to a prohibited coercive intervention, a breach of the target state's territorial inviolability or independence of state powers or, in extreme cases, even a use of force. However, the aspects of attributing the dissemination of disinformation to a state and of demonstrating a causal nexus between disinformation and effect that are necessary for international responsibility to arise have not been sufficiently addressed in the literature. This article analyses the challenges that contemporary forms of digital disinformation create for proving attribution pursuant to the customary rules of state responsibility as well as the issue of causation. In doing so, it investigates the content of the primary rules for clues pertaining to the necessary causal nexus and assesses different standards of causation employed in international and domestic law.

1 Introduction

From the outset, the ongoing COVID-19 pandemic has been accompanied by what the World Health Organization calls an 'infodemic',[1] the unprecedented surge of inaccurate information about the virus, its spread, the public response and remedies

* Hauser Post-Doctoral Global Fellow, New York University School of Law, USA; Program Leader International Cyber Law, Digital Society Institute, European School of Management and Technology Berlin, Germany; Associate Research Fellow, Geneva Academy of International Humanitarian Law and Human Rights, Switzerland. Email: henning.lahmann@esmt.org. I would like to thank Mariana Velasco-Rivera for her very helpful comments on earlier drafts of this article.

[1] World Health Organization, *Coronavirus Disease 2019: Situation Report 45*, 5 March 2020, at 2, available at www.who.int/docs/default-source/coronaviruse/situation-reports/20200305-sitrep-45-covid-19.pdf?sfvrsn=ed2ba78b_4.

and vaccinations, widely disseminated online and more often than not propagated across borders with the presumable intention to mislead target audiences in other countries.[2] Particularly in light of the fact that a considerable number of disinformation campaigns seem to have been initiated or carried out by or on behalf of governments, some scholars have renewed calls to 'shore up international law'[3] against this escalating phenomenon.[4] In 2020, the Netherlands became the first state to declare that such state-led information operations would be considered potential violations of international law.[5]

Against this background, I inquire into the ways in which states may be held internationally responsible for conducting or tolerating health-related disinformation and influence campaigns. Section 2 discusses the recent literature that considers which primary rules of international law might be engaged by such activity. I then turn to analyse two of the most intricate problems in relation to disinformation that have not received the appropriate attention in legal scholarship: the question of attribution and that of the causal relationship between disinformation and the consequences of such conduct at the stage of the breach of a primary rule – an issue notoriously underexplored in international legal scholarship.[6] As will be seen in section 3, the question of attribution in the context of the contemporary, globally connected digital information landscape presents a number of distinct challenges that deserve their own consideration separate from the otherwise widely studied realm of cyber conduct. After briefly addressing obligations of prevention that might circumvent the attribution problem in section 4, section 5 examines the role of causation in the law of state responsibility, investigates the primary rules identified in section 2 for clues concerning the required causal nexus between an act and a result and surveys possible standards of causation: 'substantial contribution', the 'NESS account' and presumed causation. These standards are subsequently discussed in detail in the context of the mechanics of the contemporary disinformation landscape.

[2] European Commission, *Tackling COVID-19 Disinformation: Getting the Facts Right*, Doc. JOIN(2020) 8 final, 10 June 2020, at 1, available at https://ec.europa.eu/info/sites/info/files/communication-tackling-covid-19-disinformation-getting-facts-right_en.pdf. This article focuses on disinformation as false or misleading information created to intentionally deceive the public. See European Commission, *Tackling Online Disinformation: A European Approach*, Doc. COM(2018) 236 final, 26 April 2018, at 3, available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0236&from=EN. Malign intent is what distinguishes dis- from misinformation.

[3] G.P. Corn, 'Coronavirus Disinformation and the Need for States to Shore Up International Law', *Lawfare* (2 April 2020), available at www.lawfareblog.com/coronavirus-disinformation-and-need-states-shore-international-law.

[4] J. Bright et al., Coronavirus Coverage by State-Backed English-Language News Sources. Understanding Chinese, Iranian, Russian and Turkish Government Media, 8 April 2020, available at https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/04/Coronavirus-Coverage-by-State-Backed-English-Language-News-Sources.pdf.

[5] *The Kingdom of the Netherlands' Response to the Pre-draft Report of the OEWG*, April 2020, at para. 18, available at https://front.un-arm.org/wp-content/uploads/2020/04/kingdom-of-the-netherlands-response-pre-draft-oewg.pdf.

[6] See Plakokefalos, 'Causation in the Law of State Responsibility and the Problem of Overdetermination: In Search of Clarity', 26 *European Journal of International Law* (*EJIL*) (2015) 471, at 473, n. 7.

2 Transboundary Health Disinformation: A Violation of International Law?

Although, as mentioned, state-led influencing activities that interfere with democratic decision-making processes in other states – together with the problem of disinformation and the increasing distortion of the global information space more generally – have received a growing amount of scholarly attention over the past half decade, there has been no consensus to date as to what rules of international law such conduct might violate.[7] Most frequently invoked in the existing literature are a broadly understood obligation to respect the sovereignty of other states[8] – provided this rather indistinct concept can be considered a standalone rule under customary law[9] – the prohibition of intervention[10] and the right to self-determination; the latter either by itself[11] or in connection with the principle of non-intervention.[12]

Disseminating false or misleading information about health issues, particularly amid a pandemic – whether to advocate dubious remedies or to cast doubt on public health measures – is of a different quality, irrespective of the question of what such conduct ultimately aims to achieve strategically.[13] It is, in the words of UN Secretary General Antonio Guterres, a 'poison that is putting even more lives at

[7] See Lahmann, 'Information Operations and the Question of Illegitimate Interference under International Law', 53 *Israel Law Review* (2020) 189.

[8] Schmitt, '"Virtual" Disenfranchisement: Cyber Election Meddling in the Grey Zones of International Law', 19 *Chicago Journal of International Law* (2018) 30.

[9] For a detailed discussion only, see M.N. Schmitt (ed.), Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Tallinn Manual 2.0) (2017), rule 4; H. Moynihan, The Application of International Law to State Cyberattacks: Sovereignty and Non-Intervention (2019), paras 30–72. A growing number of states seem to accept the conception of 'sovereignty as a rule'. For an overview, see P. Roguski, 'Application of International Law to Cyber Operations: A Comparative Analysis of States' Views', Hague Program for Cyber Norms Policy Brief (2020); but see critically Lahmann, 'On the Politics and Ideologies of the Sovereignty Discourse in Cyberspace', 32 *Duke Journal of Comparative and International Law* (forthcoming).

[10] Baade, 'Fake News and International Law', 29 *EJIL* (2019) 1357; J. Nye, *Protecting Democracy in an Era of Cyber Information War*, 13 November 2018, available at www.hoover.org/research/protecting-democracy-era-cyber-information-war.

[11] Ohlin, 'Did Russian Cyber Interference in the 2016 Election Violate International Law?', 95 *Texas Law Review* (2017) 1579.

[12] Tsagourias, 'Electoral Cyber Interference, Self-Determination and the Principle of Non-Intervention in Cyberspace', in D. Broeders and B. van den Berg (eds), *Governing Cyberspace: Behavior, Power, and Diplomacy* (2020) 45.

[13] See, e.g., EU vs Disinfo, *The Kremlin and Disinformation about Coronavirus*, 16 March 2020, available at https://euvsdisinfo.eu/the-kremlin-and-disinformation-about-coronavirus/ ('[t]he important task for any kind of message, broadcast to an international audience from pro-Kremlin outlets, is to sow discord. Pro-Kremlin disinformation outlets expose the target audience with dozens of different statements, versions, explanations, "leaks", "sensational revelations", conspiracy theories. All this aims to diminish the trust in the efforts of the health care system, the authorities, national and international institutions. Sowing panic and distrust; creating an image of an imminent collapse; suggesting a breakdown of institutions. In the examples above, one outlet claims the information about the coronavirus is exaggerated; the other – that the apocalypse is here').

risk'.[14] Nonetheless, it has been suggested that, depending on the circumstances, the principles of non-intervention and sovereignty provide an appropriate framework to deal with transboundary health disinformation.[15] With respect to the former, an act of outside influencing supposedly qualifies as prohibited intervention if it aims at subordinating the target state's sovereign will in relation to a subject matter that is part of its *domaine réservé* through coercion.[16] While there can be little doubt that a state's public health policies are part and parcel of its sovereign prerogative, it is less clear how false or misleading information about health-related issues may count as coercive conduct in this sense. Concerning disinformation that attempts to interfere in democratic decision-making processes, it has been argued that deceptive political information may inhibit the ability of the electorate to freely choose between the different options on the ballot,[17] but this is not at stake in the context of health policies – at least not directly. Accordingly, Marko Milanovic and Michael Schmitt suggest that, as long as the disinformation 'does not substantially deprive the target state of the ability to manage the epidemic', the coercion threshold is not met.[18] However, a more expansive reading of 'coercion' seems possible in the sense that it would be sufficient for the interfering activity to hamper 'the target state in relation to the exercise of its sovereign functions in some way'.[19] Crucially, a number of states appear to share such a broader understanding.[20]

Even if one is of the view that health disinformation cannot be considered coercive for the prohibition of intervention to be engaged, as mentioned, a more recent trend in international legal scholarship and among a growing number of states assumes a primary 'rule of sovereignty' that may be implicated by a state's transboundary dissemination of harmful disinformation. This development in state practice – as primarily demonstrated through official statements concerning the application of international law to states' cyber operations – notwithstanding,[21] it might be doctrinally more precise to zoom in on specific rules derived from the principle of sovereignty, such as the right to territorial inviolability or the right to independence of state powers.[22] Given

[14] See B. Chappell, 'U.N. Chief targets "Dangerous Epidemic of Misinformation" on Coronavirus', *National Public Radio* (14 April 2020), available at www.npr.org/sections/coronavirus-live-updates/2020/04/14/834287961/u-n-chief-targets-dangerous-epidemic-of-misinformation-on-coronavirus?t=1589983773870.

[15] See Milanovic and Schmitt, 'Cyber Attacks and Cyber (Mis)Information Operations during a Pandemic', 11 *Journal of National Security Law and Policy* (2020) 247.

[16] Jamnejad and Wood, 'The Principle of Non-Intervention', 22 *Leiden Journal of International Law* (2009) 345, at 348.

[17] Nye, *supra* note 10; Baade, *supra* note 10, at 1363–1364.

[18] Milanovic and Schmitt, *supra* note 15, at 269.

[19] Moynihan, *supra* note 9, para. 148.

[20] Explicitly in reference to health policies, see *Kingdom of the Netherlands' Response*, *supra* note 5, para. 18; New Zealand, The Application of International Law to State Activity in Cyberspace (2020), para. 10. Concerning the internal political system more broadly, see Australia, Non Paper: Case Studies on the Application of International Law in Cyberspace (2020), at 10; Germany, On the Application of International Law in Cyberspace (2021), at 5–6.

[21] For an overview, see Lahmann, *supra* note 9.

[22] *Ibid.*; similarly Moynihan, *supra* note 9, para. 30.

that both direct and indirect effects can lead to a violation of sovereignty,[23] an adversarial, state-led disinformation campaign may implicate one or both of the two rules. For example, if a false or misleading piece of information induced citizens of the target state to ingest a supposedly remedial, but in fact harmful, substance that results in severe illness or even death, the right to territorial inviolability would be breached. Moreover, the functioning of essential services, public safety infrastructures or other critical state capacities is protected by the right to the independence of state powers,[24] so transboundary interference by way of the dissemination of disinformation that directly or indirectly results in disruption would amount to a violation according to this view. If citizens abandon public health guidelines because they have come to believe that the virus does not in fact exist due to false narratives circulating online, thereby thwarting the official response to the pandemic, the operation intrudes on the target state's 'sovereign power to maintain public order'.[25] In both scenarios, the state carrying out the information operation would thus be responsible for wrongful conduct.

Finally, Milanovic and Schmitt go so far as to suggest that, depending on the circumstances, state-led disinformation campaigns may even 'rise to the level of a use of force'.[26] To substantiate their argument, the authors point to the emerging consensus regarding the application of international law to cyber operations and the findings of the International Group of Experts that drafted the *Tallinn Manual 2.0*, stipulating that an adversarial campaign conducted through cyberspace amounts to a use of force within the meaning of Article 2(4) of the Charter of the United Nations if it results in significant damage, injury or death to an extent that would be considered a use of force if brought about by kinetic means.[27] Consequently, they conclude that mis- or disinformation should also be qualified as force in this sense if it directly leads to sickness or death on a considerable scale.[28]

3 Attributing the Dissemination of Disinformation

The availability, ubiquity and pervasiveness of social media and other means of digital communication have enabled an ever-expanding number of different actors – news outlets, bloggers, social media influencers, YouTube broadcasters, podcasters, trolls and so on – to engage in what, at times, is apparently random and spontaneous but, at other times, is presumably calculated and coordinated dissemination of false or misleading information online that more often than not addresses audiences across borders. To hold a state responsible for such activity in case it amounts to a violation

[23] Milanovic and Schmitt, *supra* note 15, at 254.
[24] Moynihan, *supra* note 9, para. 68.
[25] *Ibid.*, para. 122.
[26] Milanovic and Schmitt, *supra* note 15, at 269.
[27] *Tallinn Manual 2.0*, *supra* note 9, rule 69.
[28] Milanovic and Schmitt, *supra* note 15, at 269.

of a rule of international law and to enable 'response action',[29] it must be attributed – with 'clear and convincing' evidence[30] – in accordance with the customary rules reflected in Articles 4–11 of the International Law Commission's (ILC) Articles on the Responsibility of States for Internationally Wrongful Acts (ARSIWA).[31] Which of these secondary rules is applicable in a given situation depends on the circumstances of the case.

For one, organs of a state, such as intelligence services, whose conduct is attributable pursuant to Article 4 of ARSIWA, may sometimes carry out information operations themselves. However, to the extent that such things are publicly known, experience from the past few years suggests that such agencies are more likely concerned with some sort of preparatory activity, such as a cyber operation aimed at obtaining compromising data that may be useful for carrying out an influencing campaign as a subsequent step.[32] To be sure, attributing such malicious cyber conduct, as analytically distinct from information operations, entails its own intricate technical and legal questions that are beyond the scope of this article.[33] More often, states will harness media organizations or other private actors to conduct disinformation campaigns against an adversary instead of going ahead themselves, which renders the issue of attribution much less straightforward. Even the activities of state media (that is, an entity that technically belongs to the state) are not by default attributable pursuant to Article 5 of ARSIWA, which considers acts by agents that are empowered to exercise elements of state authority an act of the state. It is already questionable whether the type of activity under scrutiny here – the dissemination of information via news reports or other media activity – can be described as exercising 'state authority'.[34] Ownership does not in itself allow for the piercing of the corporate veil;[35] so long as the media organization retains editorial independence, the state cannot be held responsible even if the outlet evidently confines itself to parroting the government line.[36]

[29] Tsagourias and Farrell, 'Cyber Attribution: Technical and Legal Approaches and Challenges', 31 *EJIL* (2020) 941, at 942.

[30] To establish the responsibility of a state, the standard of proof is usually considered to be one of 'clear and convincing evidence'. See H. Lahmann, *Unilateral Remedies to Cyber Operations: Self-Defence, Countermeasures, Necessity, and the Question of Attribution* (2020), at 70–79; Brunner, Dobrić and Pirker, 'Proving a State's Involvement in a Cyber-Attack: Evidentiary Standards before the ICJ', 25 *Finnish Yearbook of International Law* (2019) 75, at 81–89; but see Tsagourias and Farrell, *supra* note 29, at 958–959 (identifying a variety of different standards).

[31] International Law Commission (ILC), Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA), UN Doc. A/56/83, 3 August 2001.

[32] This is what happened prior to the 2016 US presidential election, when the Main Directorate of the General Staff of the Armed Forces of the Russian Federation (GRU) hacked into the servers of the Democratic National Committee to copy emails that were subsequently utilized for an information operation against Hillary Clinton's campaign. See *United States v. Viktor Borisovich Netyksho et al.*, Defendants, Case 1:18-cr00215-AB, District Court for the District of Columbia, 13 July 2018.

[33] See, in detail, Tsagourias and Farrell, *supra* note 29.

[34] See *ibid.*, at 953.

[35] J. Crawford, State Responsibility: The General Part (2013), at 161.

[36] See A. Toler, 'How (Not) to Report on Russian Disinformation', *Bellingcat* (15 April 2020), available at www.bellingcat.com/resources/how-tos/2020/04/15/how-not-to-report-on-russian-disinformation/.

Disinformation originating with, or disseminated by, non-state actors can only be regarded as the conduct of the state if the requirements of Article 8 of ARSIWA are met, which provides that, for attribution to occur, the person or entity must be 'in fact acting under the instructions of, or under the direction or control of, that State in carrying out the conduct'. As for the first alternative, the main question is how specific the instructions to the private actor need to be. In *Bosnian Genocide*, the International Court of Justice (ICJ) sought to clarify the notion by stating that it is not sufficient for the instructions to have been given 'generally in respect of the overall actions taken by the persons or groups of persons having committed' the unlawful acts but, rather, 'in respect of each operation'.[37] Even if this is to be interpreted in such a way as, 'where ambiguous or open-ended instructions are given, acts which are considered incidental to the task in question or conceivably within its expressed ambit may be considered attributable to the State', this will only hold true for a limited number of influence operations that pursue a fairly precise goal.[38] An example would be the manipulative social media campaign carried out ahead of the 2016 US presidential election, when, as determined in an assessment by US intelligence, the Internet Research Agency, a private organization, was 'ordered' to carry out an influence campaign by Russian president Vladimir Putin.[39] When an assignment is much broader and more vague than this, for instance by merely instructing a private entity to flood the information ecosystem in an adversarial state with contradictory or otherwise misleading narratives, without specifying any concrete objectives – which appears to be the norm rather than the exception – it is less clear whether the 'instruction' requirement of Article 8 of ARSIWA would be met under this standard. To be 'directed' within the rule's scope would require the state to actively guide the non-state actor,[40] which will likewise not be the case if the government or other state authority merely expresses indistinct desires as to the substance or goal of a disinformation campaign.

Attribution under the 'effective control' test that the ICJ has repeatedly held as the applicable standard of attributing non-state actor conduct under the 'control' prong of Article 8 of ARSIWA[41] is likely even more difficult, given that the Court deemed neither 'financing, organising, training' nor 'the planning of the whole of its operation' sufficient,[42] indicating concrete 'domination over the act' itself by the state as the decisive criterion.[43] According to recent reports, states increasingly appear to rely on

[37] Case Concerning Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro), Judgment, 26 February 2007, ICJ Reports (2007) 43, at 208.

[38] Crawford, *supra* note 35, at 145.

[39] Office of the Director of National Intelligence, Assessing Russian Activities and Intentions in Recent US Elections, Doc. ICA 2017-01D, 6 January 2017, at ii. Despite these findings, the link to the Russian government has occasionally been called into question. See, e.g., A. Maté, 'These Questions for Mueller Show Why Russiagate Was Never the Answer', *The Nation* (23 July 2019), available at www.thenation.com/article/archive/questions-mueller-russiagate/.

[40] Tsagourias and Farrell, *supra* note 29, at 954.

[41] *Bosnian Genocide*, *supra* note 37, at 210.

[42] Case Concerning Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States), Merits, Judgment, 26 June 1986, ICJ Reports (1986) 14, at 64.

[43] Tsagourias and Farrell, *supra* note 29, at 954.

so-called 'information operations for hire' or 'black public relations' companies that offer a whole range of influence activities, thus further weakening the link between the disinformation and the commissioning state and making it ever more difficult to prove state involvement, enabling the latter to always retain a degree of plausible deniability.[44] This trend does not even spare the critical matter of public trust in coronavirus vaccines.[45] As correctly pointed out by Nicholas Tsagourias and Michael Farrell in relation to the attribution of malicious conduct in the digital realm more broadly, most publicly available reports of investigations make factual claims about the relationship between state and agent that do not satisfy the legal language of Article 8 of ARSIWA.[46]

The key insight in this context is that attributing singular instances of potentially harmful conduct is inherently difficult within the contemporary digital disinformation landscape because many different actors spread false and misleading information constantly and without necessarily pursuing distinct or readily identifiable aims. It is likely that it will become only more so in the future, with new tactics emerging that further remove the state, malign intentions notwithstanding, from the incriminated conduct. Many instances of disinformation emerge spontaneously and spread across social media in an uncontrolled manner, which provides states with ample opportunity to exploit existing distortive narratives to their benefit, if only by means of a re-tweet that is, if need be, explicitly 'not an endorsement'. Even with a wealth of incriminating clues, proving state involvement in disinformation campaigns will remain uniquely challenging under the currently applicable legal framework.[47]

Mindful of these fundamental challenges, Tsagourias and Farrell consider a number of suggestions to modify the existing rules to render successful attribution more likely. Although addressing adversarial cyber conduct and not the issue of disinformation, some of them could be applied to the latter – most pertinently, shifting the required degree of 'control' pursuant to Article 8 of ARSIWA from 'effective' to the less demanding 'overall' or even 'soft control' to account for the peculiar features of the digital realm and for the fact that the state's role is often limited to ideologically or politically influencing the non-state actor.[48] The authors further ponder whether a lowering of the applicable evidentiary standard to demonstrate the existence of a sufficiently strong link between state and agent to one of 'preponderance of the evidence' would

[44] See Silverman, Lytvynenko and Kung, 'Disinformation for Hire: How a New Breed of PR Firms Is Selling Lies Online', *Buzzfeed News* (6 January 2020), available at www.buzzfeednews.com/article/craigsilverman/disinformation-for-hire-black-pr-firms; S. Vavra, 'Facebook Is Observing a "Steady Growth" in Disinformation-for-Hire Services', *Cyberscoop* (11 May 2021), available at www.cyberscoop.com/facebook-zacarias-castillo-mexico-disinformation/.

[45] J. Henley, 'Influencers Say Russia-linked PR Agency Asked Them to Disparage Pfizer Vaccine', *The Guardian* (25 May 2021), available at www.theguardian.com/media/2021/may/25/influencers-say-russia-linked-pr-agency-asked-them-to-disparage-pfizer-vaccine.

[46] Tsagourias and Farrell, *supra* note 29, at 954–955.

[47] See, e.g., B. Nimmo *et al.*, 'Secondary Infektion', *Graphika* (2020), available at https://secondaryinfektion.org/downloads/secondary-infektion-report.pdf.

[48] Tsagourias and Farrell, *supra* note 29, at 961–965.

be appropriate in the cyber context.[49] While discussing the merits and potential downsides of these proposals is out of the scope of this article, it is important to note that reconsidering the design and substance of the existing legal frameworks for the digital age generally and for the disinformation nexus specifically might become inevitable if lack of accountability for malicious conduct continues to prevail.

4 Obligations of Prevention

To some extent, the problem of attributing disinformation can be circumvented by focusing not on the state's negative obligation to refrain from such conduct but, instead, on a possible positive obligation to undertake measures to ensure that disinformation campaigns cannot be carried out by private actors from its territory or to take measures to halt such activity when ongoing. The ICJ formulated the existence of such a principle in its *Corfu Channel* decision, holding that 'it is every State's obligation not to allow knowingly its territory to be used for acts contrary to the rights of other States'.[50] In the wake of the COVID-19 global health crisis, a couple of authors have argued that this principle may apply to transboundary health disinformation.[51]

The requirement that the private actor's conduct be 'contrary to the rights of other States' has generally been interpreted as meaning that it must have amounted to an internationally wrongful act had it been carried out by a state.[52] Accepting this premise would require assessing whether the disinformation campaign would have violated a rule of international law as discussed above. The standard for the territorial state's obligation is due diligence, which implies that it is one of conduct, not of result: the state must make reasonable efforts within its capacity to act in order to prevent non-state actors from engaging in conduct against the rights of another state from its territory.[53] A further precondition to trigger the due diligence obligation, according to the ICJ, is that the state had knowledge of the unlawful conduct in question whereby constructive knowledge – that the state should have known of the private actor's behaviour – is supposedly sufficient.[54]

Whereas the existence of such a positive obligation may seem relatively straightforward when it comes to traditional malicious cyber operations carried out by private actors from the territory of the duty-bearing state, the picture is significantly complicated in the context of the transboundary dissemination of (dis)information. Even if we assume a state's positive obligation in principle, we need to carefully weigh it against the countervailing right of individuals to the freedom of expression under international human rights law. As the European Court of Human Rights (ECtHR) has

[49] *Ibid.*, at 965–966.

[50] *Corfu Channel Case (United Kingdom v. Albania)*, Merits, Judgment, 9 April 1949, ICJ Reports (1949) 4, at 22.

[51] See, e.g., Milanovic and Schmitt, *supra* note 15, at 279–282; Coco and de Souza Dias, '"Cyber Due Diligence": A Patchwork of Protective Obligations in International Law', 32 *EJIL* (2021) 1, at 23.

[52] *Tallinn Manual 2.0*, *supra* note 9, rule 6, para. 18.

[53] *Ibid.*, rule 7, para. 16.

[54] *Ibid.*, rule 6, para. 39.

determined, a positive legal duty to act cannot extend further than the state's capacity to act, which comprises not only factual capabilities but also legal constraints, such as applicable human rights of all individuals within its jurisdiction.[55] As one of the fundamental principles of any democratic society,[56] and contrary to what Singapore's Court of Appeals recently stipulated, whether the statement made by an individual is protected by the right to freedom of expression cannot depend on the statement's truth value alone.[57]

Article 19(3) of the International Covenant on Civil and Political Rights, for instance, makes the restriction of the right contingent on the need to achieve a legitimate objective.[58] Yet even potentially harmful outcomes of the dissemination of certain false or misleading information should not be considered sufficient grounds to curtail the right. While the protection of public health is explicitly listed as one such legitimate objective, it is important to note that 'disinformation' is inherently elusive and difficult to define in law,[59] a fact that considerably facilitates government overreach or the exploitation of the problem as a pretext to crack down on civil rights more broadly. To some extent, an open society must be prepared to tolerate falsehoods in its political discourse even if adverse and even harmful outcomes are to be expected. False or misleading narratives about the effects and risks of vaccines in general, an issue that long preceded the current pandemic,[60] are a case in point. Instead of suppressing such speech through blanket bans or even criminalization, state authorities should counter with disseminating 'reliable and trustworthy information' that corrects potentially harmful falsehoods.[61] States are therefore generally not under a positive obligation to act against such disinformation in case it becomes available in another state.

This general consideration in favour of free speech even in the context of transboundary health disinformation could shift if other circumstances that render an overall different calculation imperative come into play. To assess whether an interference with the freedom of expression is 'necessary in a democratic society', as Article 10(2) of the European Convention on Human Rights specifies, the ECtHR, for example, queries, *inter alia*, whether there exists a 'pressing social need' to that effect.[62]

[55] See ECtHR, *Osman v. United Kingdom*, Appl. no. 87/1997/871/1083, Judgment of 28 October 1998, para. 116.

[56] UN Human Rights Committee (UNHRC), General Comment no. 25, UN Doc. CCPR/C/21/Rev.1/Add.7, 12 July 1996, para. 25.

[57] *Online Citizen Pte Ltd v. Attorney-General and Another Appeal and Other Matters*, [2021] SGCA 96, 8 October 2021, paras 99–100. For an analysis of the judgment, see L. Schuldt, 'In Singapore's War on Fake News, the Constitution Is Not an Obstacle', *Verfassungsblog* (16 November 2021), available at https://verfassungsblog.de/fake-news-obstacle/.

[58] International Covenant on Civil and Political Rights 1966, 999 UNTS 171.

[59] D. Kaye, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression: Disease Pandemics and the Freedom of Opinion and Expression, UN Doc. A/HRC/44/49, 23 April 2020, para. 42.

[60] See Burki, 'Vaccine Misinformation and Social Media', 1 *Lancet Digital Health* (2019) e258, available at https://doi.org/10.1016/S2589-7500(19)30136-0.

[61] See Joint Declaration on Freedom of Expression and 'Fake News', Disinformation and Propaganda, Doc. FOM.GAL/3/17, 3 March 2017, principle 2(d).

[62] See, e.g., ECtHR, *The Sunday Times v. United Kingdom*, Appl. no. 6538/74, Judgment of 26 April 1979, para. 59.

A quickly unfolding global health crisis caused by a hitherto unknown pathogen – a situation entirely different in scale and urgency from 'normal' health disinformation concerning, for example, the risk of autism from the MMR vaccine[63] – could provide a 'pressing social need' in this sense. If it can additionally be shown that there in fact exists a causal relationship between the disinformation and adverse health outcomes,[64] the territorial state might then indeed be obligated to actively suppress the information.[65]

Even in situations in which such a positive duty exceptionally exists, any measures to prevent the (transboundary) dissemination of the disinformation must adhere to the principle of proportionality. This means that only the least intrusive instrument that furthers the objective may be chosen.[66] Neither the criminalization of publishing disinformation[67] nor the blanket bans of websites or other media outlets will meet this requirement.[68] This points to the general problem with the entire concept of a positive obligation to limit information: the incentives that it creates for authoritarian governments to curtail freedom of expression domestically are evident as numerous new, severely restrictive laws in various countries since the onset of the COVID-19 pandemic – ostensibly with the purpose of suppressing potentially harmful disinformation – have laid bare.[69] Thus, while it is difficult to refute the existence of such a positive obligation amid a global health crisis in principle, it should be limited to extreme and obvious cases of coordinated campaigns.

5 The Intricacies of Speech Act Causation

Even more vexed than attribution in the context of disinformation is the issue of causation. By itself, the dissemination of information does not lead to any adverse consequences. What is required is an individual that receives the information, processes it and turns it into reasons that form the basis of subsequent behaviour (for example, to ingest a toxic substance that allegedly fends off the coronavirus, to decide against wearing a mask or to not get vaccinated). While some scholars have argued that the relationship between a speech act and the recipient's subsequent conduct cannot be described in causal terms as this would contradict the free will postulate,[70] most

[63] See DeStefano and Shimabukuro, 'The MMR Vaccine and Autism', 6 *Annual Review of Virology* (2019) 585.

[64] This would require demonstrating a theory of general causation applicable to the circumstances of the case, as will be discussed in detail later in this article.

[65] Likewise Milanovic and Schmitt, *supra* note 15, at 272.

[66] UNHRC, General Comment no. 34, UN Doc. CCPR/C/GC/34, 12 September 2011, para. 34.

[67] Kaye, *supra* note 59, para. 42.

[68] UNHRC, *supra* note 68, para. 43.

[69] See Malaret and Chrobak, 'The Criminalization of COVID-19 Clicks and Conspiracies', *DFRLab* (13 May 2020), available at https://medium.com/dfrlab/op-ed-the-criminalization-of-covid-19-clicks-and-conspiracies-3af077f5a7e7.

[70] In a Hegelian tradition, see, e.g., Puppe, 'Der objektive Tatbestand der Anstiftung', *Goltdammer's Archiv für Strafrecht* (1984) 101, at 108–110.

contemporary legal theorists follow Gilbert Ryle's refutation of the Cartesian idea that mind and body are separate entities,[71] conceiving mental processes as physical processes that are therefore subject to causal laws,[72] which nonetheless does not necessarily preclude the idea of a free will.[73] However, even if the laws of causation apply to mental processes, how a certain speech act directed at an individual influences that individual's behaviour is much more difficult to predict, measure and generalize than 'purely physical' phenomena, such as what will happen when water is heated to a hundred degrees Celsius.[74]

In this pragmatic sense, Richard Wilson has proposed the notion of 'mental causation', as opposed to 'physical causation', to describe the connection between communicative acts, the minds of the recipients and the decisions based on the communicative link that lead to observable consequences.[75] This distinction is not foreign to international law: in the process of drafting the Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, the ILC sought to exclude activities that involve 'human psychology' due to the difficulty to establish a causal relationship in such areas with certainty.[76] To avoid confusion as to the precise meaning of the notion of 'mental causation', which is also used to describe what happens inside a person's mind,[77] I use the term 'speech act causation' in what follows to denote the causal relationship between an instance of the dissemination of disinformation and the consequences of that act.

A *Causation in the Law of State Responsibility*

The customary law of state responsibility as reflected in ARSIWA is mostly silent on the issue of causation. It is only made explicit at the stage of reparation: Article 31 of ARSIWA states that '[t]he responsible State is under an obligation to make full reparation for the injury caused by the internationally wrongful act' and that '[i]njury includes any damage … caused by the internationally wrongful act of a State'. At the stage of breach that is in relation to the question whether a state is responsible for

71 G. Ryle, *The Concept of Mind* (1949).

72 Morse, 'The Moral Metaphysics of Causation and Results', 88 *California Law Review* (2000) 879, at 890.

73 Wright, 'The NESS Account of Natural Causation: A Response to Criticisms', in R. Goldberg (ed.), *Perspectives on Causation* (2011) 285, at 313.

74 See Wright and Puppe, 'Causation: Linguistic, Philosophical, Legal and Economic', 91 *Chicago-Kent Law Review* (2016) 461, at 487, n. 109.

75 R.A. Wilson, Incitement on Trial. Prosecuting International Speech Crimes (2017), at 31.

76 ILC, Articles on Prevention of Transboundary Harm from Hazardous Activities, UN Doc. A/CN.4/SER.A/1987/Add.1 (Part 2) (1987), para. 155. The limitation to transboundary harm resulting from activities 'through a chain of physical events' is reflected in Article 1 of the 2001 Draft Articles, which clarifies that only 'physical consequences' are covered, although the phrasing is confusing (para. 126). Arguably, this implies that the Draft Articles are not applicable to consequences stemming from transboundary disinformation, unlike previously suggested in the literature. See, e.g., Sander, 'Democracy under the Influence: Paradigms of State Responsibility for Cyber Influence Operations on Elections', 18 *Chinese Journal of International Law* (2019) 1, at 49–52.

77 See, e.g., Siffert, 'What Does It Mean to Be a Mechanism? Stephen Morse, Non-reductivism, and Mental Causation', 11 *Criminal Law and Philosophy* (2017) 143.

violating a primary rule of international law,[78] the causation issue is merely implicit in Article 16, which presupposes the existence of some form of causal nexus between an assisting state's conduct and the breach of a primary rule for that state to be internationally responsible.[79]

Additionally, Article 12 of ARSIWA only provides that '[t]here is a breach of an international obligation by a State when an act of that State is not in conformity with what is required of it by that obligation'. The ILC explains its silence on causation by stating that this is to be dealt with by the content of the primary rule in question, although these rules usually do not include details of the required causal nexus explicitly either.[80] Therefore, it is 'for courts and practitioners to develop appropriate tests for different types of obligations'.[81] Causation at the stage of breach is rarely discussed in international jurisprudence for the simple reason that it is usually obvious that the adverse consequences that gave rise to the breach were the result of the respondent state's conduct or omission.[82] But, as indicated above, this is not the case when it comes to the mechanics of disinformation. Accordingly, it is necessary to examine the requirement and nature of a causal nexus in regard to the primary rules identified in section 2 as possibly engaged by the transboundary dissemination of disinformation.

B *Disinformation: The Necessary Causal Nexus at the Stage of Breach*

As stipulated by the ILC, the investigated primary rules should indicate whether a causal link between a state's implicated conduct (or, more precisely, the conduct of the person or entity that is attributable to the state in accordance with Articles 4–11 of ARSIWA) and some consequence is required for the rule to be engaged and, if so, what is the nature of that link. I now turn to analyse these rules. Concerning the principle of non-intervention, some scholars have correctly pointed out that a coercive action does not need to be successful for the rule to be violated.[83] At first glance, this might suggest that the rule does not require the existence of a causal nexus. However, Harriet Moynihan and the majority of the Group of Experts that drafted the *Tallinn Manual 2.0* contend that, for the *actus reus* of intervention to be satisfied, there still must have been some kind of coercive effect, which can only be qualified as such if there was a causal relationship between the coercive behaviour and the attempted deprivation of the target state's authority *vis-à-vis* its *domaine réservé*.[84] For example, the *Tallinn Manual* stipulates that, for a threat made by one state against another to violate the non-intervention principle, it does not need to successfully compel the target

[78] On this distinction, see Stern, 'The Obligation to Make Reparation', in J. Crawford, A. Pellet and S. Olleson (eds), *The Law of International Responsibility* (2010) 563, at 569; Plakokefalos, *supra* note 6, at 474.

[79] See H.P. Aust, Complicity and the Law of State Responsibility (2011), at 212.

[80] Stern, *supra* note 78, at 569–570; Plakokefalos, *supra* note 6, at 474 (who mentions Article 139 of Part XI of the Convention on the Law of the Sea as one exception). UN Convention on the Law of the Sea 1982, 1833 UNTS 3.

[81] Shelton, 'Righting Wrongs: Reparations in the Articles on State Responsibility', 96 *American Journal of International Law* (*AJIL*) (2002) 833, at 846.

[82] See Plakokefalos, *supra* note 6, at 481; Aust, *supra* note 79, at 211.

[83] Moynihan, *supra* note 9, paras 101–102; *Tallinn Manual 2.0*, *supra* note 9, at 322.

[84] Moynihan, *supra* note 9, para. 104; *Tallinn Manual 2.0*, *supra* note 9, at 320.

state to take some action against its will, but it must at least be sufficiently 'coercive in nature', which would seem to exclude obviously empty or implausible threats.[85]

An influence operation as coercion by indirect means – which the ICJ has recognized as a possible form of unlawful intervention[86] – would require showing a causal link between the information disseminated by the state and an act of the recipients of that information that exerted the necessary coercive effect on the target state. Other authors similarly assume the need for at least some causal relationship between the intervening act and an effect by zooming in on the question whether the conduct at hand led to a 'deprivation of free choice'[87] or otherwise affected the target state's 'control' over a matter under its *domaine réservé*.[88] The official statements by the few states that have formulated a position concerning disinformation and intervention support the interpretation that some coercive effect must have arisen for the rule to be engaged.[89] Accordingly, the spreading of false or misleading information that aims at disrupting the target state's public health efforts during a pandemic that no one notices or takes up will entirely fail to produce any coercive effect and thus would not constitute a violation of the principle of non-intervention. Perhaps in part due to the difficulty in establishing a causal link between an interfering act and coercive effects, there might be an emerging trend to instead make the existence of prohibited intervention turn on other factors such as (malign) intent, the mode of conduct (deceptive instead of open transboundary dissemination of information)[90] or the mere possibility of harmful outcomes.[91] Whether this marks a wider shift towards a broader understanding of

[85] *Tallinn Manual 2.0*, *supra* note 9, at 322–323; it is worth adding for clarification that, while threats can be made explicitly or implicitly, it is difficult to conceive the covert dissemination of disinformation as a 'threat' in this sense, as this mode of conduct aims at manipulating the target audience's behaviour without the manipulation being noticed, which is very different from the effects of a 'threat', be it explicit or implicit.

[86] *Nicaragua*, supra note 42, para. 205.

[87] Kilovaty, 'The Elephant in the Room: Coercion', 113 *AJIL Unbound* (2019) 87, at 90; see also Damrosch, 'Politics across Borders: Nonintervention and Nonforcible Influence over Domestic Affairs', 83 *AJIL* (1989) 1, at 49 (emphasizing the need for an interfering conduct to produce an 'effect' on the target population for the non-intervention principle to be engaged).

[88] N. Tsagourias, 'Electoral Cyber Interference, Self-Determination and the Principle of Non-Intervention in Cyberspace', *EJIL:Talk!* (26 August 2019), available at www.ejiltalk.org/electoral-cyber-interference-self-determination-and-the-principle-of-non-intervention-in-cyberspace/.

[89] See *Kingdom of the Netherlands' Response*, *supra* note 5, para. 18 ('information operations that *intervene with*'); New Zealand, *supra* note 20, para. 10 ('cyber disinformation operation that *significantly undermines* a state's public health efforts'); Germany, *supra* note 20, at 5–6 ('disinformation … *significantly impeding* the orderly conduct of an election') (emphases added).

[90] See K. Berzina and E. Soula, 'Conceptualizing Foreign Interference in Europe', *Alliance for Securing Democracies* (18 March 2020), at 2ff, available at https://securingdemocracy.gmfus.org/wp-content/uploads/2020/03/Conceptualizing-Foreign-Interference-in-Europe.pdf; Council of the European Union, *Complementary Efforts to Enhance Resilience and Counter Hybrid Threats – Council Conclusions*, Doc. 14972/19, 10 December 2019; US Cybersecurity and Infrastructure Security Agency, *Foreign Interference*, available at www.cisa.gov/publication/foreign-interference; Australian Department of Home Affairs, *Countering Foreign Interference*, available at https://homeaffairs.gov.au/about-us/our-portfolios/national-security/countering-foreign-interference.

[91] In this sense, see French Ministry of Defence, International Law Applied to Operations in Cyberspace (2019), at 7 ('interference which causes *or may cause harm* to France's political, economic, social and cultural system, may constitute a violation of the principle of non-intervention'; emphasis added).

intervention in the future remains to be seen. So far, Lassa Oppenheim's declaration that '[i]nterference pure and simple is not intervention' still rings true.[92]

The same rationale applies to a possible violation of the rights to territorial inviolability or the independence of state powers. For disinformation to amount to a breach of the rule, the dissemination must have resulted in tangible territorial impacts or an actual usurpation of inherently governmental functions.[93] As to the use of force, while the word 'use' may merely indicate the prohibition of a certain type of conduct, the notion of 'force' implies the need for some kind of effect in close relationship to a cause.[94] This argument is supported by *Oil Platforms*, where the ICJ came closest to engaging in an actual causal analysis at the stage of breach in its jurisprudence on the use of force.[95] The USA was compelled to show not only that one of its vessels had been hit by a missile but also where and when it had been launched as well as its flight path to establish a clear link between the origin and the result. Considering disinformation, it seems far-fetched to construct the transboundary dissemination of a potentially harmful piece of false information (for example, the recommendation to ingest methanol to avoid an infection with COVID-19) that no individual takes up as a use of force within the ambit of Article 2(4) of the UN Charter. To be sure, the rule prohibits not only the actual use of force but also the threat of it. But, as argued above, even if we consider the possibility of merely implying the threat to use force, which is sufficient for the rule to be breached,[96] it is difficult from a semantic perspective to conceive the concealed dissemination of disinformation that aims at subtly manipulating the behaviour of the target audience as a 'threat' in this sense.

Finally, the positive obligation to prevent harmful health disinformation is a 'true obligation of prevention' in light of Article 14(3) of ARSIWA, meaning that there is no breach of the rule unless the harm actually materializes.[97] Considerations regarding the freedom of expression alone render it infeasible to construct the rule as one that might be violated by a state's failure to exercise due diligence, without any actual adverse consequences from disinformation. This implies the need to establish a causal link between the information and the harmful behaviour of non-state actors that the state had the duty to prevent.

92 R. Jennings and A. Watts (eds), *Oppenheim's International Law*, vol. 1: *Peace* (9th edn, 2008), at 432.

93 Hollis, 'The Influence of War; The War for Influence', 32 *Temple International and Comparative Law Journal* (2018) 31, at 42. Concerning the right to self-determination, which was identified above as another rule probably engaged by an influence campaign targeting democratic decision-making processes, the causal issues would be the same, as the rule presupposes an actual impact on the voting public. See *ibid.*, at 43.

94 *Tallinn Manual 2.0*, *supra* note 9, at 334.

95 *Case Concerning Oil Platforms (Iran v. United States)*, Judgment, 6 November 2003, ICJ Reports (2003) 161, at 189–190. It should be noted that the Court did not explicitly undertake a causal analysis; rather, it mixed aspects of causation and attribution without referring to a clear conceptual framework.

96 *Tallinn Manual 2.0*, *supra* note 9, at 338–339; D.B. Hollis and T. van Benthem, 'What Would Happen If States Started Looking at Cyber Operations as a "Threat" to Use Force?', *Lawfare* (30 March 2021), available at www.lawfareblog.com/what-would-happen-if-states-started-looking-cyber-operations-threat-use-force.

97 Crawford, *supra* note 35, at 227–228.

C *Standards of Causation*

The brief survey of the pertinent primary rules above reveals that (i) each requires an observable effect to have materialized as a result of the state's conduct or omission and (ii), in the case of disinformation, the relationship between the conduct (dissemination) and the effect is by definition indirect as it can only occur through the influencing of individuals. To prove this connection, it is thus necessary to examine what kind of standard of causation to prove this connection is appropriate and applicable to the identified primary rules of international law. While the previous section shows that these rules allow for inferences regarding the need of a causal link, they are silent on how this link is to be established. Generally speaking, international law does not contain specific rules to tackle questions of proving causation.[98] Given that it is not inherently imperative that the same causal standard applies to each of the primary rules,[99] it therefore makes sense to look for guidance in other fields of law, specifically because 'issues of causality have less to do with the individual legal field or system than with the idea of law itself'.[100] Ilias Plakokefalos suggests taking insights from tort law due to its 'strong structural similarities' to the law of state responsibility,[101] and Anthony Aust, dealing with questions of state complicity, refers to international criminal law for its wealth of doctrinal clarification on the issue.[102] As both fields have to tackle the intricacies of speech act causation, doctrinal solutions derived from either seem suitable to consider for the matter at hand.

To properly understand the differences between the standards discussed in the following section, it is important to clarify two critical analytical distinctions from the outset. The first is between what may be called 'natural' (or 'factual') causation and 'legal' causation, more appropriately described as the 'scope of responsibility'. While the former concerns the natural chain of events leading to a certain event, the latter refers to the evaluative consideration of an event whereby factors that are part of the causal chain are disregarded in the final assessment of responsibility for legal reasons.[103] The construct, which is recognized in international law,[104] applies both at the stage of breach and at the stage of reparation.[105] The second is the distinction between correlation and general and specific causation. Correlation describes the occurrence of two variables – for example, an instance of disinformation and a harmful event – within reasonably close temporal and contextual proximity. In itself, it is unable to reveal anything with certainty about the causal relationship between them.[106] While correlation may indicate causation and can thus be utilized to reduce complexity and

98 Plakokefalos, *supra* note 6, at 476.
99 Aust, *supra* note 79, at 218.
100 *Ibid.*, at 214.
101 Plakokefalos, *supra* note 6, at 475–476.
102 Aust, *supra* note 79, at 214.
103 Crawford, *supra* note 35, at 492; Morse, *supra* note 72, at 891.
104 Plakokefalos, *supra* note 6, at 475.
105 *Ibid.*, at 492.
106 Barrowman, 'Correlation, Causation, and Confusion', 43 *The New Atlantis* (2014) 23, at 30.

improve the efficiency of the causal analysis,[107] inferring a causal link from observed correlation alone is fallacious.[108] As H.L.A. Hart and Tony Honoré have noted, 'not all events which follow each other in invariable sequence are causally related'.[109] The concept of general causation denotes scientifically validated generalizations of causality that allow for the conclusion that a certain observable phenomenon or activity (for example, smoking or the spreading of false narratives about a virus online) is capable of leading to adverse outcomes (for example, lung cancer or the exacerbation of a public health crisis).[110] Specific causation, on the other hand, applies a causal generalization to an observed temporal and contextual chain of events and is thus necessary to decide a singular case.[111]

1 Substantial Contribution

The 'substantial contribution' test has been consistently applied by the international criminal tribunals for Rwanda and the former Yugoslavia in relation to accomplice culpability,[112] and it is also frequently invoked by domestic courts in common law jurisdictions[113] and mentioned in the ILC's commentary in regard to liability for aid and assistance pursuant to Article 16 of ARSIWA.[114] Most significantly, both the International Criminal Tribunal for Rwanda (ICTR) and the International Criminal Tribunal for the former Yugoslavia (ICTY) have employed the standard in cases of instigation – a mode of liability[115] that holds a perpetrator accountable who has prompted

[107] For example, 'excess mortality' – the number of deaths over a given period of time above that to be expected under 'normal' conditions – can be useful to assess the impact of a health crisis, but it is not suitable to substitute a proper cause-of-death investigation. It is, however, an appropriate first step to determine what to look for. See Geijingting *et al*., 'Correlation Analysis and Causal Analysis in the Era of Big Data', 563 *IOP Conference Series: Materials Science and Engineering* (2019) 1, at 5.

[108] The *post hoc ergo propter hoc* fallacy. See C.T. Bergstrom and J.D. West, *Calling Bullshit: The Art of Scepticism in a Data-Driven World* (2020), at 56–68; Wilson, *supra* note 75, at 126–131 (who argues that the fallacy played out during the International Court for the former Yugoslavia [ICTY] *Šešelj* trial).

[109] H.L.A. Hart and T. Honoré, *Causation in the Law* (2nd edn, 1985), at 15.

[110] General causation can usually be explained as probabilistic cause (that is, A increases the chance of B in a causal manner). See Bergstrom and West, *supra* note 108, at 72–73.

[111] See generally S. Haack, *Evidence Matters: Science, Proof, and Truth in the Law* (2014), at 269. An example: Proposition 1: 'Smoking causes lung cancer' states a relationship of general causation. Proposition 2: 'Person A (a) smokes and (b) has lung cancer, *therefore* (c) their smoking caused the cancer' is one of specific causation. The truth value of proposition (2) does not by itself follow from (1) even if (1) is true on the basis of scientific evidence. Despite a strong suggestion of (2) being true as well, this claim requires further corroboration, which involves the consideration of other possible causes – for example, A working in a factory that emits toxic fumes.

[112] G.S. Gordon, Atrocity Speech Law: Foundation, Fragmentation, Fruition (2017), at 344.

[113] Wright and Puppe, *supra* note 7474, at 480–481.

[114] The ARSIWA Commentary holds that the aid or assistance must have 'contributed significantly' to the other state's wrongful act. 'State Responsibility, General Commentary' (ARSIWA Commentary) 2(2) *ILC Yearbook* (2001) 31, Art. 16, para. 5 (this is materially no different from 'substantial contribution'). See Crawford, *supra* note 35, at 403.

[115] Coco, 'Instigation', in J. de Hemptinne (ed.), *Modes of Liability in International Criminal Law* (2019) 257, at 257.

another person to commit an offence[116] and that therefore requires the establishment of a causal link between instigator and instigated.[117] Instigation is different from the crime of incitement, which, as an inchoate offence, does not require the target audience to act upon the inciting speech act.[118] This renders it superfluous to show a causal nexus.[119] The fact that instigation – giving rise to liability for a speech act – depends upon a proven causal link to another person's subsequent behaviour creates a strong resemblance to the purported effects of disinformation: mentally inducing the speech act's recipient to engage in a harmful act.

Both the ICTR and the ICTY determined that, regarding the defendant's act of influencing in cases involving instigation, it was 'not necessary to demonstrate that the crime would not have occurred without the accused involvement'[120] in the sense of a strict *conditio sine qua non* or 'but for' test.[121] Instead, the tribunals declared it sufficient to establish that the instigation constituted a 'clear'[122] or 'substantially contributing'[123] factor. However, the standard has been criticized for a lack of clarity as to what it actually means in practice. Aside from the statement that the necessary causal relationship cannot be affirmed in cases of an *omnimodo facturus* (that is, if the actual perpetrator had already definitely decided to commit the crime irrespective of the instigating speech act),[124] there is no concrete guidance in the tribunals' case law. Moreover, Gregory Gordon has observed that, especially in later decisions, the test was interpreted so strictly as to essentially amount to the 'but for' standard of causality.[125] For instance, in its *Šešelj* decision, the ICTY did not find the defendant criminally liable for instigation despite the fact that it could be proven that he had called for the expulsion of Croats and that the latter had been expelled by individuals who had listened to his instigating speech, asserting instead that 'the Prosecution was not able to marshal evidence that this speech would have been *at the root of* the departure of the Croats'.[126] Apart from its vagueness, some authors have argued that the standard is doctrinally

[116] Judgment, *Brđanin* (IT-99-36-T), Trials Chamber, 1 September 2004, para. 269.

[117] E. van Sliedregt, Individual Criminal Responsibility in International Law (2012), at 104.

[118] Judgment, *Akayesu* (ICTR-96-4-T), Trials Chamber, 2 September 1998, para. 562; Ohlin, 'Incitement and Conspiracy to Commit Genocide', in P. Gaeta (ed.), *The UN Genocide Convention: A Commentary* (2009) 186, at 193.

[119] See Gordon, *supra* note 112, at 246. This was misunderstood by Agbor, who argued that neither instigation nor incitement require a causal nexus. See Agbor, 'The Substantial Contribution Requirement: The Unfortunate Outcome of an Illogical Construction and Incorrect Understanding of Article 6(1) of the Statute of the ICTR', 12 *International Criminal Law Review* (2012) 155.

[120] Judgment, *Kvočka* (IT-98-30/I-T), Trials Chamber, 2 November 2001, para. 252.

[121] Judgment, *Orić* (IT-03-68-T), Trials Chamber, 30 June 2006, para. 274.

[122] Judgment, *Blaškić* (IT-95-14/I-T), Trials Chamber, 3 March 2000, para. 270; *Kvočka*, *supra* note 120, para. 252; *Brđanin*, *supra* note 116, para. 269.

[123] Judgment, *Ndindabahizi* (ICTR-2001-71-I), Trials Chamber, 15 July 2004, para. 463; Judgment, *Kordić and Cerkez* (IT-95-14/II-A), Appeals Chamber, 17 December 2004, para. 27; *Orić*, *supra* note 121, para. 274; Judgment, *Nahimana et al.* (ICTR-99-52-A), Appeals Chamber, 28 November 2007, para. 501.

[124] *Orić*, *supra* note 121121, para. 274.

[125] See Gordon, *supra* note 112, at 250 (with reference to the *Šešelj* case before the ICTY).

[126] Judgment, *Šešelj* (IT-03-67-T), Trials Chamber, 31 March 2016, para. 333 (emphasis added).

flawed by impermissibly conflating aspects of natural causation and evaluative questions regarding the scope of responsibility.[127]

2 The NESS Account of Causation

Building on Hart and Honoré's work[128] and partly in response to the confusion of the different stages of a proper causal analysis, Richard Wright has devised the 'necessary element of a sufficient set' (NESS) account of causality.[129] The NESS account states that a 'particular condition is a cause of (contributed to) a specific result if and only if it was a necessary element of a set of antecedent actual conditions that was sufficient for the occurrence of the result'.[130] In cases of speech act causation, establishing that the reception of some information contributed to a specific decision to behave in a certain way (that is, was causal), according to this standard, it is sufficient to demonstrate that the information was in fact considered by the recipient and counted positively in favour of the decision to act.[131] If such positive consideration can be proven,[132] it does not affect the causal relationship if the decision itself was overdetermined, for example because the recipient agent had taken into account additional sources of information that also contributed to the decision or if she would have acted in the incriminated manner anyway.[133] Consider the *Šešelj* decision described earlier. According to the NESS account, there is little doubt that the defendant's speech was a causal factor for the expulsion of the Croats.

It is important to note that the standard itself only concerns the issue of natural causation. The overall analysis of the *actus reus* is still subject to considerations regarding the scope of responsibility, which might be precluded, for example, if a piece of information was considered but played only a remote, minor part in the decision to act. Even if natural causation can be demonstrated, it remains possible that 'strongly necessary intervening conditions' ultimately negate legal responsibility for the outcome.[134]

[127] Plakokefalos, *supra* note 6, at 475; Wright and Puppe, *supra* note 74, at 480–481.

[128] Hart and Honoré, *supra* note 109, at 112–113.

[129] See Wright and Puppe, *supra* note 74, at 464.

[130] Wright, 'Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts', 73 *Iowa Law Review* (1988) 1001, at 1019.

[131] *Ibid.*, at 1037; Wright and Puppe, *supra* note 74, at 487, n. 109; similarly J. Gardner, *Offences and Defences: Selected Essays in the Philosophy of Criminal Law* (2007), at 70–71.

[132] One innovative approach to proving a link between speech act and the recipient's subsequent behaviour is the identification of 'mental fingerprints', recurring words, phrases or expressions that demonstrate a likely relationship. See Dojčinović, 'Word Scene Investigations: Toward a Cognitive Linguistic Approach to the Criminal Analysis of Open Source Evidence in War Crimes Cases', in P. Dojčinović (ed.), *Propaganda, War Crimes Trials and International Law* (2012) 71. For a possible example, see Judgment, *Nahimana et al.* (ICTR-99-52-T), Trials Chamber, 3 December 2003, para. 476.

[133] See Puppe, *supra* note 70, at 110.

[134] Wright and Puppe, *supra* note 74, at 502; Morse, *supra* note 72, at 891. For some factors that may preclude the scope of responsibility in international law, see Crawford, *supra* note 35, at 492.

3 Presumed Causation

A further standard applied in domestic tort law is that of 'presumed causation'. In cases where it is difficult to demonstrate with reasonable certainty the adverse consequences of the wrongful act (for example, libel), the standard is sometimes invoked to relieve the plaintiff of the need to prove that the harm was in fact caused by that act.[135] Presumed causation effectively amounts to a reversal of the burden of proof in that the defendant needs to present evidence that establishes the contrary.[136] Notably, the standard is usually applied at the stage of reparation, not at the stage of breach. It rests on the premise that the existence of an unlawful act has already been established. Accordingly, any subsequently emerging negative consequences that could reasonably be connected to the act (for example, financial loss due to reputational damage after a libellous act) are then attributed to it. Investigating the effects of the dissemination of transboundary disinformation requires establishing unlawfulness in the first place.

In relation to disinformation, presumed causation is conceivable in a strong and a weak variant. For the strong presumption, causation would be considered proven based on observed correlation alone. That means it would only be necessary to show that there is false or misleading information with a certain content, the dissemination of which is attributable to a state, and that there are temporally subsequent adverse consequences that roughly 'fit' the original message (for example, information that slanders the candidate of a national election who subsequently loses). The strong presumption will effectively amount to a prohibition of the interfering activity in and of itself because the lack of a causal nexus will be virtually impossible to demonstrate for a state accused of having spread the false narratives even if it can show that identical falsehoods originated from different sources not attributable to the state. In relation to the adverse consequences, the standard would approximate strict liability, which is doctrinally questionable given that the concept usually only concerns the question of fault and negligence and does not encompass causality.[137]

In its weaker form, for the presumption to be triggered, one would not only need to demonstrate the interfering conduct, attribution and some correlated adverse outcome but also put forth a theory of general causation that credibly and verifiably establishes that the conduct (a certain type of disinformation) is reasonably likely to result in the observed consequences (for example, inducing voters to

[135] See Anderson, 'Reputation, Compensation, and Proof', 25 *William and Mary Law Review* (1984) 747, at 764.

[136] Wright and Puppe, *supra* note 74, at 481. In common law, the presumption is sometimes taken as preventing the defendant from even introducing facts that support alternative theories of causation. See Anderson, *supra* note 135, at 764–765.

[137] See, e.g., ILC Commentary to the Draft Principles on the Allocation of Loss in the Case of Transboundary Harm Arising Out of Hazardous Activities (2006), Principle 4, para. 16 ('[s]trict liability may alleviate the burden that victims may otherwise have in proving fault of the operator, but it does not eliminate the difficulties involved in establishing the necessary causal connection of the damage to the source of the activity').

abandon a viable candidate). This weaker variant of presumed causation has some precedent in domestic legislation regarding environmental harm. The German Environmental Liability Act, for example, instead of requiring the establishment of a causal link between the observed damage and an emitter, focuses on the latter's capability of producing that damage.[138] While this standard is tempting for contexts in which it is inherently difficult to prove causation, the reluctance of most states to accept the inclusion of interactions involving 'human psychology' (such as economic or social matters) in the scope of the Draft Articles on Prevention of Transboundary Harm from Hazardous Activities, due to misgivings regarding proof of causation, raises doubts about their willingness to embrace it in the context of disinformation.[139] Furthermore, the ICJ's strict insistence on a complete presentation of the chain of events in *Oil Platforms* suggests that the Court is not likely to apply presumptions at least in cases that concern international responsibility for the use of force.[140]

D *Proving Speech Act Causation*

Having introduced three standards of causation that may be suitable in the context at hand, this section, with the help of insights from the cognitive and social sciences, takes an in-depth look at how the contemporary (online) information landscape actually functions and at how the effects of targeted campaigns or spontaneous instances of false or misleading information unfold and are observable and measurable. Despite a quickly growing body of literature on the phenomenon, scholarly claims remain at a rather unspecific level – such as that disinformation erodes 'the very foundation of open societies'[141] and that, irrespective of the question of impact, adversarial actors certainly try to negatively affect target societies.[142] Only a more thorough investigation into the causal mechanics of disinformation will enable us to assess what legal constructs might be feasible to hold states accountable for the dissemination or toleration of potentially harmful health disinformation. One of the main problems with the current treatment of transboundary disinformation is that many studies, if addressing the problem of causation at all, conflate the notion of correlation and causation or fail to distinguish between general and specific causation. To reiterate, depending on the applied standard, proving causation in a given instance requires evidence in relation to one, two or all three of these aspects. The outcome of the analysis will vary accordingly.

[138] Wacker-Theodorakopoulos and Kreienbaum, 'Environmental Damage and the Question of Liability', 27 *Intereconomics* (1992) 157, at 159–160. Section 6(1)1 of the Gesetz über die Umwelthaftung (Environmental Liability Act), 10 December 1990, BGBl. I, 2634, provides: 'If an installation is likely to cause the damage that occurred on the basis of the given facts of the individual case, it is presumed that the damage was caused by this installation'.

[139] ILC, *supra* note 76, paras 152–155.

[140] *Oil Platforms*, *supra* note 95, at 189–190.

[141] T. Rid, Active Measures: The Secret History of Disinformation and Political Warfare (2020), at 11.

[142] Y. Benkler, R. Faris and H. Roberts, *Network Propaganda* (2018), at 267.

1 Disinformation as Audience Manipulation

Social and cognitive science scholars have explained the mechanics of disinformation as the manipulation, as opposed to the persuasion, of the recipient of a communicative act. Whereas 'persuasion' can be understood as 'a process of appealing… to reason',[143] manipulation is described as 'directly influencing someone's beliefs, desires, or emotions such that she falls short of ideals for belief, desire, or emotion in ways typically not in her self-interest or likely not in her self-interest in the present context'.[144] Disinformation manipulates through the induction of misperceptions in the recipient's mind – that is, 'factual beliefs that are false or contradict the best available evidence in the public domain'.[145] Misperceptions curtail a person's autonomy in that they interfere with their control over their own process of reasoning, whereas persuasion leaves this process and thus autonomy intact.[146] In *Nahimana et al.*, the ICTR Trial Chamber explicitly zoomed in on this distinction when examining broadcasts by Radio-Télévision Libre des Mille Collines (RTML) that claimed that Rwanda's Tutsi population was disproportionately wealthier than the Hutu. If correct, the tribunal held, the assertion could be qualified as 'an effort to disseminate information to the public on inequities of social concern'. Conversely, if the statement was false, it 'might be considered an attempt to *manipulate public opinion* and generate unfounded hostility toward and resentment of the Tutsi population'.[147] An information's falseness can relate to different aspects, such as the truth value of the content of a piece of information, the identity of the speaker or both; many disinformation campaigns seek to manipulate their audience with factually correct information by concealing its true source.[148] The digital transformation has provided a range of tools to bring about misperceptions alongside the ability to generate fake and inauthentic accounts on social media – for example, bots to artificially amplify a message's reach to create a false sense of significance or micro-targeting algorithms to adjust content to the preconceptions of certain subgroups of society.[149]

For the purpose of establishing a causal link, the standards of causation discussed above appraise the issue of manipulation and persuasion differently. This may be illustrated by the following example. Assume *The Guardian*, a news organization based in London, publishes a detailed report on structurally racist practices at the police department in Minneapolis, including cover-ups of unjustified fatal shootings of people of colour. The exclusive story leads to outrage in the community and to intense protests, in the course of which some participants cause considerable

[143] Strauss, 'Persuasion, Autonomy, and Freedom of Expression', 91 *Columbia Law Review* (1991) 334, at 335.

[144] Barnhill, 'What Is Manipulation?', in C. Coons and M. Weber (eds), *Manipulation: Theory and Practice* (2014) 51, at 52.

[145] Flynn, Nyhan and Reifler, 'The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs about Politics', 38 *Advances in Political Psychology* (2017) 127, at 128.

[146] Strauss, *supra* note 143, at 354–355; see Kaye, *supra* note 59, para. 60.

[147] *Nahimana et al.*, Trials Chamber, *supra* note 132, para. 470 (emphasis added).

[148] Rid, *supra* note 141, at 10.

[149] See European Commission, *Tackling COVID-19 Disinformation*, *supra* note 2, at 4.

property damage in the city. Responding to the protesters' intensifying pressure, the city administration announces drastic policy changes, and the district attorney decides to indict three police officers who are accused of having been involved in wrongful deaths. The question is: did *The Guardian* piece contribute to the outcome in a causal manner?

For adherents of the 'substantial contribution' standard, the answer must be no because the aspect of manipulation directly impacts the causation analysis. Merely reporting correct facts, the dispatch did not create any misperceptions. Retaining their autonomy, the recipients' own judgment led them to engage in the harmful behaviour, and this judgment thus superseded the contribution of the speech act.[150] In that sense, the causal connection to the information was disrupted by the recipients' free decision as an intervening cause.[151] This is how the international criminal tribunals have approached the issue.[152] Apart from the doctrinal reservations against this theory discussed above, it can be hard to draw a clear line between persuasion and manipulation. Danny Scoccia has suggested to conceive them not as binaries but, rather, as 'two opposite ends of a continuum'.[153] Reliably determining an instance of manipulation in a non-arbitrary way is therefore inherently difficult, especially amid a confusing situation like an unfolding health crisis that sees frequent shifts in what is considered 'the best available evidence in the public domain', destabilizing the epistemic consensus against which information can be assessed.[154] The standard may have difficulty in accounting for such subtleties.

The NESS account, by contrast, has no trouble in assuming a causal link. It only needs to be demonstrated that at least some of the protesters had in fact read *The Guardian* piece and that it had counted positively towards their decision to act. Whether or not the decision was truly autonomous or manipulated is immaterial; of relevance are only the recipients' inner motives, not how these had come about.[155] It also would not matter if the piece had merely provided one reason among many and if the protests had happened either way. But, crucially, the analysis of natural causation is only the first step. To assess whether the United Kingdom may have *prima facie* breached its obligation to prevent such an adverse outcome on the territory of the USA, it is decisive to additionally consider the scope of responsibility.[156] Here, factors such as the report's truthfulness, the publisher's lack of malign intent and the open and transparent way of dissemination should be taken into account. The fact that the recipients acted on the basis of an autonomous decision may even count as 'strongly necessary intervening condition' that, while not affecting causality, results in a denial

[150] Scanlon, 'A Theory of Freedom of Expression', 1 *Philosophy and Public Affairs* (1972) 204, at 212.

[151] Wilson, *supra* note 75, at 148, 169.

[152] See, e.g., *Nahimana et al.*, Trials Chamber, *supra* note 132, para. 470; *Šešelj*, *supra* note 126, para. 333.

[153] See Scoccia, 'Can Liberals Support a Ban on Violent Pornography?', 106 *Ethics* (1996) 776, at 784–785.

[154] See Simpson and Srinivasan, 'No Platforming', in J. Lackey (ed.), *Academic Freedom* (2018) 186, at 191–192.

[155] Puppe, *supra* note 70, at 109.

[156] Plakokefalos, *supra* note 6, at 478.

of legal responsibility.[157] This latter consideration would also introduce the question of manipulation into the assessment, with the same pitfalls.

A causal analysis, according to the strong standard of 'presumed causation', would reach the same conclusion, as it would suffice to show correlation between the report and the subsequent events in Minneapolis. The 'weak presumption' test would only be satisfied with additional scientifically valid research that establishes a theory of general causation connecting this type of reporting to the observed consequences. Either way, the issue of manipulation is irrelevant for the causal analysis. However, as the 'presumed causation' standard is usually applied at the stage of reparation, when the existence of an unlawful act has already been determined, there needs to be some instrument to adjust for unreasonable results if applied to the stage of breach. Therefore, to avoid overreach, evaluative considerations regarding the scope of responsibility should complement the application of this standard.

2 Manipulation of Behaviour and Manipulation of Mental States

The question whether a piece of information was manipulative seems to provide a suitable factor for the legal assessment either in the causal analysis itself or in relation to the scope of responsibility. This insight is important for the context of disinformation in light of a further distinction introduced by Anne Barnhill: manipulation targeting behaviour and manipulation targeting mental states, such as attitudes or beliefs.[158] The former 'aims to change what someone immediately decides or does',[159] for instance, when a radio broadcast names specific individuals who are supposedly members of an enemy armed group, as was the case with some RTML transmissions in Rwanda in April 1994.[160] The latter attempts to influence the recipients' existing belief system without trying to directly impact their actions. One frequently occurring example are protracted disinformation campaigns that aim to gradually raise hostility towards a minority, as happened in Rwanda[161] or, more recently, against the Rohingya in Myanmar,[162] where the report of the 2018 Independent International Fact-Finding Mission found 'a carefully crafted hate campaign' that 'developed a negative perception of Muslims among the broad population in Myanmar'.[163]

The two modes of manipulation will often be combined to enhance their effects by targeting the majority group's attitudes over a longer period of time, 'infecting the

[157] Wright and Puppe, *supra* note 74, at 502.

[158] Barnhill, *supra* note 144, at 57.

[159] *Ibid.*

[160] See *Nahimana et al.*, Trials Chamber, *supra* note 132, paras. 477–478.

[161] *Ibid.*, para. 470.

[162] P. Mozur, 'A Genocide Incited on Facebook, with Posts from Myanmar's Military', *New York Times* (15 October 2018), available at www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html.

[163] Report on the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar (Rohingya Report), UN Doc. A/74/486, 9 October 2019, para. 696.

mind'[164] with dehumanizing narratives about the persecuted minority until it merely takes a minor trigger to instigate physical violence.[165] This method was on display in Rwanda, when a protracted campaign of ethnic stereotyping culminated first in explicit calls for the extermination of the Tutsi population and then in the calling out of individuals belonging to the ethnic group, some of whom were subsequently murdered.[166] In Myanmar, the fact-finding report concluded that long-running hateful 'discourse created a conducive environment for the 2012 and 2013 anti-Muslim violence'.[167] When it was supplemented with a false report that two Muslim teashop owners had raped their Buddhist maid, anti-Muslim pogroms ensued.[168]

Since the start of the COVID-19 pandemic, we have seen instances of both types of manipulation. Whereas claims on social media that ingesting methanol or drinking milk helps to fend off the virus belong to the first category,[169] the narrative that the virus does not exist, that vaccines have no preventive function but are part of the agenda for a 'new world order'[170] and will be used to inject nano chips to control the population are instances of the second.[171] Another method of manipulative disinformation targeting mental states is the strategy to disseminate a high number of incoherent narratives that aim to confuse the audience, leading it to lose trust in institutional authority and eventually faith in the concept of truth itself.[172]

3 The Mechanics of Contemporary Disinformation

There is reason to assume that the vast majority of contemporary disinformation, especially if disseminated by state actors, is best characterized as such slowly unfolding

[164] In the words of the judges at the war crimes trial against Nazi propagandist Julius Streicher, '[i]n his speeches and articles, week after week, month after month, he infected the German mind with the virus of anti-Semitism and incited the German people to active persecution'. *The Nuremberg Trial*, 6 FRD 69, 161–163 (International Military Tribunal, 1946).

[165] On the effects of dehumanizing narratives, see Guillard and Harris, 'The Neuroscience of Dehumanization and Its Implications for Political Violence', in P. Dojčinović (ed.), *Propaganda and International Law: From Cognition to Criminality* (2020) 201.

[166] *Nahimana et al.*, Trials Chamber, *supra* note 132, para. 949. On this, see Straus, 'What Is the Relationship between Hate Radio and Violence? Rethinking Rwanda's "Radio Machete"', 35 *Politics and Society* (2007) 609, at 613.

[167] Rohingya Report, *supra* note 163, para. 696.

[168] Justice Trust, *Hidden Hands behind Communal Violence in Myanmar: Case Study of the Mandalay Riots* (2015), at 20–21, available at www.burmalibrary.org/docs21/Justice_Trust-2015-03-Hidden_Hands-en-to-rev1-red.pdf.

[169] EU vs Disinfo, EEAS Special Report Update: Short Assessment of Narratives and Disinformation Around the COVID-19/Coronavirus Pandemic (Updated 2–22 April), 24 April 2020, available at https://euvsdisinfo.eu/eeas-special-report-update-2-22-april/.

[170] EU vs Disinfo, *Disinfo: Vaccines Don't Heal; Their Production Is Part of the Agenda for a New World Order*, 24 March 2020, available at https://euvsdisinfo.eu/report/vaccines-dont-heal-their-production-is-part-of-the-agenda-for-a-new-world-order/.

[171] EU vs Disinfo, *Vaccine Hesitancy and Pro-Kremlin Opportunism*, 16 April 2020, available at https://euvsdisinfo.eu/vaccine-hesitancy-and-pro-kremlin-opportunism/.

[172] See US Department of State, *GEC Special Report: Pillars of Russia's Disinformation and Propaganda Ecosystem* (2020), at 5, available at www.state.gov/wp-content/uploads/2020/08/Pillars-of-Russia%E2%80%99s-Disinformation-and-Propaganda-Ecosystem_08-04-20.pdf; EU vs Disinfo, *The Kremlin and Disinformation About Coronavirus*, 16 March 2020, available at https://euvsdisinfo.eu/the-kremlin-and-disinformation-about-coronavirus/.

attacks against the 'liberal epistemic order',[173] primarily targeting attitudes and beliefs instead of attempting to reach more tangible and immediate goals, such as influencing the outcome of an election.[174] At least one state has come forward to declare such long-term interference a possible violation of the prohibition of intervention.[175] But the diffuse causal nexus created by such conduct presents a serious problem to establish causation according to the standards introduced above.[176] Even under the assumption that adverse consequences of this type of disinformation can be observed reliably, the slow and gradual undermining of the recipients' existing value system, by definition, implies that there is not one or a clearly defined number of speech acts that have amounted to a substantial contribution to the adverse outcome.[177] As held by the ICTR Appeals Chamber, 'the longer the lapse of time' between the speech act and the harm, 'the greater the possibility that other events might be the real cause… and that [the speech act] might not have substantially contributed to it'.[178] It will rarely be possible to rule out that the recipient might have retained agency, disrupting the causal chain.[179]

The NESS account is better suited to deal with cases of overdetermination,[180] deeming any single factor sufficient that was a – instead of the – cause for the outcome. But it nonetheless demands 'concrete evidence of the actual conditions in a specific situation',[181] which means that it must be substantiated that identifiable individuals in fact considered some specific information that had an impact on their attitudes and beliefs in a way that resulted in observable adverse effects – for example, ignoring public health guidelines amid a pandemic or electing non-democratic parties or candidates. To do this, demonstrating the measurable degree of exposure to certain disinformation or the generated user engagement alone is generally not sufficient.[182]

[173] Rid, *supra* note 141, at 10–11.

[174] See J. Watts, 'Whose Truth? Sovereignty, Disinformation, and Winning the Battle of Trust', *Atlantic Council* (2018), at 4.

[175] Germany, *supra* note 20, at 6 ('cyber activities targeting elections may be comparable in scale effect to coercion if they aim at *and result in* a substantive disturbance or even permanent change of the political system of the targeted State, i.e. *by significantly eroding trust* in a State's political organs and processes', emphasis added).

[176] Hellner, 'Causality and Causation in Law', 40 *Scandinavian Studies in Law* (2000) 111, at 114 (who employs the term 'diffuse causation' to describe mental causation more generally).

[177] See A. Applebaum, 'History Will Judge the Complicit', *The Atlantic* (July/August 2020), available at www.theatlantic.com/magazine/archive/2020/07/trumps-collaborators/612250/.

[178] *Nahimana et al.*, Appeals Chamber, *supra* note 123, para. 513; see also the cautious conclusions of the Rohingya Report, *supra* note 163, paras 1325–1326; generally sceptical Straus, *supra* note 166; Carver, 'Broadcasting and Political Transition: Rwanda and Beyond', in R. Fardon and G. Furniss (eds), *African Broadcast Cultures: Radio in Transition* (2000) 188, at 190.

[179] In this direction, see Wilson, *supra* note 75, at 174–175; Straus, *supra* note 166, at 615; Scoccia, *supra* note 153, at 779–780.

[180] See Plakokefalos, *supra* note 6.

[181] Wright and Puppe, *supra* note 74, at 492.

[182] See, e.g., H. Au, J. Bright and P.N. Howard, 'Social Media Junk News: Undermining Lockdown Consensus and Consent', *Oxford Internet Institute* (26 May 2020), available at https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/06/ComProp-Coronavirus-Misinformation-Weekly-Briefing-26-05-2020.pdf; EU vs Disinfo, *Throwing Coronavirus Disinfo at the Wall to See What Sticks*, 2 April 2020, available at https://euvsdisinfo.eu/throwing-coronavirus-disinfo-at-the-wall-to-see-what-sticks/.

At least, it must be shown who was actually in touch with the content and with what intensity and duration.[183] If this is successful and leads to the establishment of a causal relationship, the consideration of the scope of responsibility might take into account factors such as the recipient's confirmation bias, though by how much is certainly up for discussion. Cognitive science research suggests that attitudes can likely only be influenced if a new piece of (dis)information connects to something that is already part of the recipient's belief system.[184] Because of this, individuals will mostly seek out information that reinforces already held beliefs in order to integrate it into a consistent worldview[185] or interpret the information according to existing biases and preconceptions.[186] Such directionally motivated reasoning has no bearing on the causal chain according to the NESS account, but one might wonder whether the contribution of a singular piece of disinformation to an adverse outcome at some point down the line, even if disseminated by a state with malign intent, does not become too remote and marginal to be considered within the scope of responsibility.[187] To be sure, such a conclusion entirely depends on the specific circumstances of the case and certainly does not preclude responsibility if the state continuously flooded the target state's information space with false narratives over an extended period. If actual consideration by recipients can be proven, the standard is thus suitable at least in relation to targeted influence campaigns.

The benefit of applying the standard of weak presumptive causation is that it does not require evidence in relation to the attitudes of concrete individuals in a concrete situation. Concerning the demonstration of theories of general causation in relation to the effects of disinformation, a growing amount of social science research strongly suggests that exposure to false or misleading narratives – most significantly, conspiracy theories – decreases trust in science and negatively affects social behaviour, including the willingness to get vaccinated or to reduce one's carbon footprint.[188]

[183] See Mironko, 'The Effect of RTML's Rhetoric of Ethnic Hatred in Rural Rwanda', in A. Thompson (ed.), *The Media and the Rwandan Genocide* (2007) 125.

[184] G.S. Jowett and V. O'Donnell, *Propaganda and Persuasion* (5th edn, 2012), at 34.

[185] See Flynn et al., *supra* note 145, at 132; Barnhill, *supra* note 144, at 55. This was already pointed out by H. Arendt, *The Origins of Totalitarianism* (2nd edn, 1958), at 351 ('[w]hat convinces masses are not facts, and not even invented facts, but only the consistency of the system of which they are presumably part').

[186] On this aspect, see Reasons of Judge Fremr, Decision on Defence Applications for Judgments of Acquittal, *William Samoei Ruto and Joshua Arap Sang* (ICC-01/09-01/11), Trial Chamber V(a), 5 April 2016, para. 141.

[187] Note that recent studies show that many recipients are aware of the falseness of a piece of information online but intentionally contribute to disseminating it further. See Madrid-Morales *et al.*, 'Motivations for Sharing Misinformation: A Comparative Study in Six Sub-Saharan African Countries', 15 *International Journal of Communication* (2021) 1200; A. Chadwick and C. Vaccari, *News Sharing on UK Social Media: Misinformation, Disinformation, and Correction* (2019).

[188] See, e.g., Jolley and Douglas, 'The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions', 9 *PLOS One* (2014), available at www.ncbi.nlm.nih.gov/pmc/articles/PMC3930676/; Jolley and Douglas, 'The Social Consequences of Conspiracism: Exposure to Conspiracy Theories Decreases Intentions to Engage in Politics and to Reduce One's Carbon Footprint', 105 *British Journal of Psychology* (2014) 35; K. Müller and C. Schwarz, *Fanning the Flames of Hate: Social Media and Hate Crime* (2017), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972; van der Linden, 'The Conspiracy-Effect: Exposure to Conspiracy Theories (About Global Warming) Decreases Pro-Social Behavior and Science Acceptance', 87 *Personality and Individual Differences* (2015) 171.

To be sure, sceptics have noted that the causal relationship between speech acts and harmful behaviour 'remains dismayingly opaque, and doggedly resistant to empirical testing'.[189] It is certainly advisable to remain cautious even if the results of a survey seem to reveal a causal link between observed attitudes or behaviour and some existing online narratives. Too often, claims about the impact of certain disinformation campaigns are asserted rather than in any way demonstrated with scientifically valid means.[190] Supposedly abductive reasoning is sometimes actually just a preference for the most convenient or readily available explanation.[191] But the progress of social science research in this field allows for the conclusion that this standard can be used to circumvent some of the problems of proving specific causation without completely neglecting the laws of natural causation.

Finally, another phenomenon points to the serious pitfalls of the strong form of presumed causation: it is possible that the recipient does not believe a piece of disinformation at all but chooses to accept it for the sole reason that it aligns with her general attitude towards a contentious topic, so the false narrative can be used to reduce the social costs of holding such a belief[192] or even to rationalize harmful behaviour after the fact.[193] Such a case will show a strong correlation between disinformation and harm, but the causal chain is in fact inverted.[194] If applied correctly, neither the NESS account nor a weak presumption (if based on valid science) could affirm a causal link in such a constellation, but the strong form likely would.

In sum, the 'substantial contribution' standard, aside from being problematic for doctrinal reasons, is under-inclusive when it comes to speech act causation as it sets the threshold too high with its focus on the question whether the information subdued the recipient's autonomy through manipulation, at least as interpreted by the international criminal tribunals in their later jurisprudence on instigation.[195] The strong presumption standard, on the other hand, suffers from overinclusion and overstretches the principles of natural causation. Assuming causation following an

[189] S. Cottee, 'Can Facebook Really Drive Violence?', *The Atlantic* (9 September 2018), available at www.theatlantic.com/international/archive/2018/09/facebook-violence-germany/569608/.

[190] See Benkler, Faris and Roberts, *supra* note 142, at 267 ('[d]oing this critically important work creates a strong bias to assume that the hard-won successful observations of intervention are a sign of large impact and threat. However, trying hard, as the Russians clearly are, does not equal actual success in affecting the outcomes or attitudes of a society at large').

[191] See Haack, *supra* note 111, at 235 ('[b]ut loose talk of "inference to the best explanation" disguises the fact that what presently seems like the most plausible explanation may not really be so – indeed, may not really be an explanation at all. We may not know all the potential causes of D, or even which other candidate-explanations we would be wise to investigate').

[192] See Flynn et al., *supra* note 145, at 138.

[193] Cottee, *supra* note 189.

[194] On reverse causality, see Blankshain and Stigler, 'Applying Method to Madness: A User's Guide to Causal Inference in Policy Analysis', 3 *Texas National Security Review* (2020), available at https://tnsr.org/2020/07/applying-method-to-madness-a-users-guide-to-causal-inference-in-policy-analysis/.

[195] Gordon, *supra* note 112, at 343–344 (who advocates for keeping the standard but to stop interpreting it in a way that effectively amounts to 'but for' causality. However, the author provides little guidance for an application of the standard in practice).

observed correlation may be tempting in the context of disinformation, as many false or misleading narratives spreading via social media appear to perfectly match undesired or harmful outcomes. But given the abundance of disinformation present in today's online media ecosystem, such inferences are easily confounded by selection bias that distorts the data set used for the analysis.[196] This leaves the NESS account and the weak standard of presumed causation. As a general theory of causation, the former is the most persuasive and precise and certainly suitable to solve the causation puzzle in cases of behaviour manipulation. But, considering the need to demonstrate the actual mental states of identified individuals, it might struggle to establish a causal nexus when a disinformation campaign targets attitudes that only very slowly, if at all,[197] turn into observable effects. As verifiable social science research can provide the basis to establish valid theories of general causation,[198] the weak presumption standard may be employed as an instrument in these contexts. Both standards preserve the possibility to adjust the result of the causal analysis by considering evaluative criteria that affect the scope of responsibility and thus deny satisfaction of the *actus reus* in a given situation.

As mentioned, there is no inherent reason why the same standard must apply to each of the potentially engaged primary rules. One might suggest that a weak presumption based on a validated theory of general causality is appropriate for assuming a breach of the principle of non-intervention but that, for a violation of a use of force, a stricter standard of causation, requiring evidence for the concrete chain of events, is necessary – as insinuated by the ICJ in *Oil Platforms*.[199] The lower the standard as to what aspects of a causal nexus must be proven, the more the primary rules will come to resemble pure prohibitions of certain activities that merely create the risk of adverse consequences.[200] As this concerns the transboundary dissemination of information, such a prohibition of interference would then need to be able to distinguish between lawful and unlawful activities, perhaps by zooming in on factors like intent or mode of conduct (open and transparent as opposed to covert and deceptive). However, as argued above, evidence of past practice suggests that this is not how states have understood the pertinent primary rules until now.

6 Concluding Observations

The proliferation of digital disinformation and state-led influence campaigns, whether directed at democratic decision-making processes or public health measures directly

[196] On this, see Blankshain and Stigler, *supra* note 194.

[197] See, e.g., Fazio, 'Multiple Processes by Which Attitudes Guide Behavior: The Mode Model as an Integrative Framework', 23 *Advances in Experimental Social Psychology* (1990) 75.

[198] See, e.g., Bastick, 'Would You Notice If Fake News Changed Your Behavior? An Experiment on the Unconscious Effects of Disinformation', 116 *Computers in Human Behavior* (2021) 106633.

[199] *Oil Platforms*, *supra* note 95, at 189–190.

[200] Note that this only concerns the stage of the breach. If an affected state sought reparations from the culpable state, the former would still need to demonstrate that the activity caused the damage. See ARSIWA, *supra* note 31, Art. 31.

or at existing fissures within societies to induce an erosion of trust more broadly, is a serious and growing concern. Yet, as shown in this article, to hold a state responsible for such activity under current international law, it is necessary to attribute the conduct pursuant to Articles 4–11 of ARSIWA and to establish a causal link between the disinformation and some observable result with clear and convincing evidence, which is ever more difficult in the contemporary disinformation environment, where the issues of attribution and causation enable states to retain plausible deniability and thus to evade international responsibility.

Against this backdrop, this article has offered an analysis of the attribution and causation puzzles in the context of disinformation to examine what doctrinal constructs might contribute to holding states accountable for disseminating or tolerating harmful transboundary disinformation. In this regard, it has been shown that, especially when it comes to speech act causation, scholarship cannot afford to ignore the epistemological insights of the social and cognitive sciences. The alternative bears significant risks of overreaction both in relation to the purportedly responsible state, possibly leading to a further escalation, and domestically, where the perception of ubiquitous threats from disinformation increases the likelihood of severe curtailments of the freedom of expression, the fundamental principle without which the very order we are trying to preserve cannot possibly function.