Human Rights Law Review, 2021, 00, 1–27
doi: 10.1093/hrlr/ngaa060
Article

Regulatory Responses to 'Fake News' and Freedom of Expression: Normative and Empirical Evaluation

Rebecca K Helm* and Hitoshi Nasu†

ABSTRACT

National authorities have responded with different regulatory solutions in attempts to minimise the adverse impact of fake news and associated information disorder. This article reviews three different regulatory approaches that have emerged in recent years—information correction, content removal or blocking, and criminal sanctions—and critically evaluates their normative compliance with the applicable rules of international human rights law and their likely effectiveness based on an evidence-based psychological analysis. It identifies, albeit counter intuitively, criminal sanction as an effective regulatory response that can be justified when it is carefully tailored in a way that addresses legitimate interests to be protected.

KEYWORDS: fake news, social media, freedom of expression, psychological bias, regulatory response

1. INTRODUCTION

Early proponents of the internet imagined an information utopia, in which information freely and easily shared would yield tremendous benefits to society.[1] However, the widespread exploitability of information on the internet generally, and social media more specifically, has allowed so called 'fake news' to impact individuals' perception of domestic and international affairs. 'Fake news' has been defined as 'fabricated information that mimics news media content in form but not in organizational process or intent.'[2] It encompasses misinformation (false or misleading information) and disinformation (false or misleading information disseminated with the specific purpose of deceiving people)—two types of information disorder.

In response to the rise of fake news through social media, national authorities have responded with different regulatory solutions in attempts to minimise or eliminate

* Senior Lecturer in Law, University of Exeter, United Kingdom.
† Professor of International Law, University of Exeter, United Kingdom.

1 See, for example, Negroponte, *Being Digital* (1995) at 158.

2 Lazer et al, 'The Science of Fake News' (2018) 359 *Science* 1094.

the adverse impact of information disorder. The implementation of these measures is fraught with difficulties, with earlier studies raising concerns about their compatibility with freedom of expression and speech.[3] Traditionally, in some parts of the world at least, freedom of speech has been treasured with the 'marketplace of ideas' metaphor, which Justice Oliver Wendell Holmes Jr articulated in his cherished dissenting judgment in *Abrams v United States*.[4]

However, increased polarization and algorithmically dictated content dissemination and consumption make this metaphor less appropriate in today's society, where truth may not be emerging from such a 'marketplace.'[5] Due to the distorting impact of fake news for democratic decision-making processes, harmful consequences may well be perceived to outweigh the benefit of free speech for society and democratic processes.[6] Normative considerations must therefore be carefully weighed against the need to ensure that regulatory responses effectively combat fake news. There are critical gaps in literature where rigorous analysis is needed for precarious balancing between normative compliance and psychological effectiveness in the crafting of the regulatory response to fake news.

This article addresses this precarious balancing by critically evaluating different regulatory approaches in terms of their normative compliance with the applicable rules of international human rights law, and their effectiveness to achieve the regulatory goal according to an evidence-based psychological analysis. To that end, it outlines the criteria that regulatory regimes should aim to meet both empirically (Section 2) and normatively (Section 3) and evaluates three different regulatory approaches based on these criteria (Section 4). This article concludes with the finding that some level of restriction on freedom of expression is inevitable due to the need to discourage the creation and distribution of fake news, rather than just preventing its spread.

In particular, this article identifies, albeit counter-intuitively, criminal sanction as an effective regulatory response. Contrary to broad normative claims worshipping freedom of expression, careful analysis of normative requirements under international law suggests that criminal sanction can be justified when it is tailored in a way that specifically, and with sufficient precision, addresses legitimate interests to be protected with varying degrees of safeguard required under each jurisdiction against abuse, including an opportunity to contest allegations of falsity. However, it is cautioned that the introduction of interventionist measures necessarily involves social cost due to the

3 See, for example, Milanovic, 'Viral Misinformation and the Freedom of Expression', *EJIL Talk!*, 13 April 2020, Parts I-III, available at: www.ejiltalk.org [last accessed 2 December 2020]; Marsden, Meyer and Brown, 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?' (2020) 36 *Computers Law & Security Review* 105373; Manzi, 'Managing the Misinformation Marketplace: The First Amendment and the Fight Against Fake News' (2019) 87 *Fordham Law Review* 2623; Katsirea, '"Fake News": Reconsidering the Value of Untruthful Expression in the Face of Regulatory Uncertainty' (2018) 10 *Journal of Media Law* 159; Richter, 'Fake News and Freedom of the Media' (2019) 8 *Journal of International Media & Entertainment Law* 1; Wilson and Umar, 'The Effect of Fake News on Nigeria's Democracy within the Premise of Freedom of Expression' (2019) 4 *Journal of Telecommunication Study* 6.

4 250 U.S. 616 at 630 (1919).

5 See, for example, Lombardi, 'The Illusion of a "Marketplace of Ideas" and the Right to Truth' (2019) 3(1) American Affairs 198; Napoli, 'What if More Speech is No Longer the Answer' (2018) 70 *Federal Communications Law Journal* 55.

6 *Chaplinsky v New Hampshire*, 315 U.S. 568 at 572 (1942). See also Craufurd Smith, 'Fake News and Democratic Legitimacy: Lessons for the United Kingdom' (2019) 11 *Journal of Media Law* 52 at 63–4.

chilling effect they have on the socially beneficial free flow of information. The extent to which social benefits of free flow of information are perceived to be outweighed by the public interest in the removal of the societal harms generated by fake news may vary depending on how much social cost each society is prepared to accept.

2. PROBLEMS OF INFORMATION DISORDER AND PSYCHOLOGICAL MECHANISMS

The spread of fake news, facilitated easily through social media, has the potential to distort public opinions and affect policy-making processes, especially in democratic societies.[7] The societal impact of this problem has been illustrated during the recent COVID-19 pandemic, where a variety of false claims have been widely circulated. These have included claims about the origins of coronavirus (for example, claims that the US military introduced the virus to Wuhan),[8] and many false health claims.[9] This led to the Vice President of the European Commission for Values and Transparency describing a Coronavirus 'infodemic,' in which disinformation has harmed the health of citizens, negatively impacted the economy, and undermined the response of public authorities.[10] Public health impacts can also be seen in other contexts; for example, fake news linking the measles, mumps, and rubella (MMR) vaccination to autism has caused a drop in childhood immunization rates for all vaccines.[11]

The problem of information disorder is also critical to democracy, including the maintenance of national security and public order. The dissemination of fake news through social media has been reported as a deliberate means of foreign interference with democratic processes in numerous jurisdictions, most notably during the 2016 US Presidential election.[12] It has also been found to incite violence.[13] The spread of

7 European Commission, *Tackling Online Disinformation: A European Approach*, COM(2018) 236 final, 26 April 2018 at 2; Kajimoto and Stanley (eds), *Information Disorder in Asia: Overview of Misinformation Ecosystem in India, Indonesia, Japan, the Philippines, Singapore, South Korea, and Taiwan* (2018).

8 Molter and Webster, 'Coronavirus Conspiracy Claims: What's Behind a Chinese Diplomat's COVID-19 Misdirection', *Internet Observatory*, 31 March 2020, available at: cyber.fsi.stanford.edu/io/news/china-covid19-origin-narrative [last accessed 2 December 2020].

9 Budoo, 'Controls to Manage Fake News in Africa Are Affecting Freedom of Expression', *The Conversation*, 11 May 2020, available at: theconversation.com/controls-to-manage-fake-news-in-africa-are-affecting-freedom-of-expression-137808 [last accessed 2 December 2020]; Howard et al, 'The COVID-19 "Infodemic": What Does the Misinformation Landscape Look Like and How Can We Respond?', Oxford Internet Institute, 15 April 2020, available at www.oii.ox.ac.uk/blog/the-covid-19-infodemic-what-does-the-misinformation-landscape-look-like-and-how-can-we-respond/ [last accessed 2 December 2020].

10 Jourová, Vice President of the European Commission for Values and Transparency, Speech in response to Disinformation around COVID-19, Brussels, 10 June 2020, available at: ec.europa.eu/commission/presscorner/detail/en/SPEECH_20_1033 [last accessed 2 December 2020].

11 See, for example, Carrieri, Madio and Principe, 'Vaccine Hesitancy and Fake News: Quasi-experimental Evidence from Italy' (2019) 28 *Health Economics* 1377.

12 See, for example, *Report of the Select Committee on Intelligence, United States Senate, on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election*, 116th Congress 1st Session, Report 116-XX, 8 October 2019, vol 2, available at: www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf [last accessed 2 December 2020]; UK House of Commons, Digital, Culture, Media and Sport Committee, *Disinformation and 'Fake News': Final Report*, HC1791, 18 February 2019, ch 6, available at: publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf [last accessed 2 December 2020]; Intelligence and Security Committee of Parliament, *Russia*, HC632, 21 July 2020 at paras 27–28, available at: isc.independent.gov.uk/news-archive/21july2020 [last accessed 2 December 2020].

13 Adegoke and BBC Africa Eye, 'Like. Share. Kill. Nigerian Police Say False Information on Facebook Is Killing People', *BBC News*, 13 November 2018, available at: www.bbc.co.uk/news/resources/idt-sh/nige

fake news does not necessarily cause direct damage; rather, it tends to affect a more diffused interest, such as public order or the integrity of democratic processes.[14] The ability to exploit online communication also adds further to the modern tools of information warfare as the means of foreign interference and sabotage.[15] Indeed, in its Resolution 2217 adopted on 26 April 2018, the Council of Europe recognised the widespread exploitability of disinformation on social media as a hybrid threat 'intended to undermine security, public order and peaceful democratic processes.'[16]

The harmful consequences of fake news necessitate effective regulatory responses to reduce or eliminate its adverse impacts on the orderly society or political processes. An examination of likely effectiveness must be based on an understanding of the psychological mechanisms that facilitate the spread of fake news, and their probable responsiveness to regulatory intervention. There is a large body of research conducted in psychology and communications regarding these psychological mechanisms. This body of research highlights the potential for polarising impact and strong persistence of fake news once it has been created.

The combined effect of two well-documented psychological biases is key to understanding the psychological mechanisms that facilitate belief in fake news, and also facilitate its spread and persistence. The first is confirmation bias. This refers to the motivation to seek out information that confirms existing beliefs, expectations or hypotheses (hereinafter 'beliefs') and the tendency to interpret information in line with these beliefs. [17] This means that once people have an existing belief, they will have a bias towards searching for and believing information that conforms with it.

The second is motivated cognition and information processing. This refers to the inclination to seek out and credit information supportive of self-defining values and attitudes (in other words, information consistent with cultural outlook). [18] Research also suggests an influence of what those receiving information 'want' to be true.[19] Thus, people are more likely to seek out and believe information that is in line with the beliefs

ria_fake_news [last accessed 12 August 2020]; 'Anti-Rumour Campaigner Lynched in Tripura', *The Hindu*, 29 June 2018, available at: www.thehindu.com/news/national/other-states/campaigner-against-rumours-lynched-in-tripura/article24294471.ece [last accessed 2 December 2020].

14 *Guide to Guarantee Freedom of Expression Regarding Deliberate Disinformation in Electoral Contexts*, OAS Doc OEA/Ser.D/XV.22 (2019) at 21.

15 See, for example, Guandagno and Guttieri, 'Fake News and Information Warfare: An Examination of the Political and Psychological Processes from the Digital Sphere to the Real World' in Chiluwa and Samoilenko (eds), *Handbook of Research on Deception, Fake News, and Misinformation Online* (2019) 167.

16 Council of Europe Parliamentary Assembly, Resolution 2217, 26 April 2018, Legal Challenges related to Hybrid War and Human Rights Obligations at para 4. See also European Commission, supra n 7 at 2.

17 See, for example, Flynn, Nyhan and Reifler, 'The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics' (2017) 38(S1) *Advances in Political Psychology* 127; Westerwick, Kleinnman and Knobloch-Westerwick, 'Turn a Blind Eye If You Care: Impacts of Attitude Consistency, Importance, and Credibility on Seeking of Political Information and Implications for Attitudes' (2013) 63 *Journal of Communication* 432.

18 Gollust et al, 'Controversy Undermines Support for State Mandates on the Human Papillomavirus Vaccine' (2010) 29 *Health Affairs* 2014; Kahan et al, 'Who Fears the HPV Vaccine, Who Doesn't, and Why? An Experimental Study of the Mechanisms of Cultural Cognition' (2010) 34 *Law and Human Behavior* 501; Kahan, Jenkins-Smith and Braman, 'Cultural Cognition of Scientific Consensus' (2011) 14 *Journal of Risk Research* 147.

19 Bastardi, Uhlmann and Ross 'Wishful Thinking: Belief, Desire, and the Motivated Evaluation of Scientific Evidence' (2011) 22 *Psychological Science* 731

of their in-groups, such as their political groups. The resulting bias can be demonstrated by looking at the politically polarised nature of misinformation endorsement. For example, in the US context, research has shown that Republicans are more likely to endorse the myth that former President Obama was not born in the USA, while Democrats are more likely to believe the myth that Bush administration officials were complicit in the 9/11 terrorist attacks.[20]

Importantly, the psychological biases described above have also been shown to influence the perceived credibility of a source of information, specifically through a disposition to consider a source more credible where its content adheres to existing beliefs or is consistent with cultural outlook (and vice versa).[21] Importantly, experts in a field are not universally seen as being credible, particularly following the rise of what has been termed 'anti-intellectualist' or 'anti-science' thought.[22] Experimental work examining expert opinion relating to scientific and political issues has shown that individuals are significantly less likely to rate an expert with elite academic credentials as 'trustworthy and knowledgeable' when they adopt the position (on climate change, nuclear waste disposal, or handgun regulation) that is opposed to the individual's cultural outlooks (and vice versa).[23] Instead, the perceived trustworthiness has been shown to be more influential than expertise in persuading people to change their views following receipt of incorrect information.[24]

This tendency to downgrade ratings of credentials that are not belief or value consistent is part of a wider phenomenon known as 'disconfirmation bias,' where information that is not consistent with pre-existing beliefs or values is judged more critically in order to discount it.[25] The tendency to find a source more trustworthy where information is consistent with pre-existing beliefs or values, and less trustworthy when it is not, creates the potential for both endorsement of belief and outlook consistent information from non-credible sources (through over-rating the credibility of those sources), and the rejection of belief and outlook inconsistent corrective information from credible sources (through under-rating the credibility of those sources). Regulatory solutions must account for these psychological effects of biases to ensure effective reduction or elimination of adverse impacts of fake news.

20 Nylan, 'Why the "Death Panel" Myth Wouldn't Die: Misinformation in the Health Care Reform Debate' (2010) 8(1) *The Forum* 1 at 3.

21 Fragale and Heath, 'Evolving Information Credentials: The (Mis) Attribution of Believable Facts to Credible Sources' (2004) 30 *Personality and Social Psychology Bulletin* 225; Kahan, Jenkins-Smith and Braman, supra n 18.

22 See, in the US context, Motta, 'The Dynamics and Implications of Anti-Intellectualism in the United States' (2017) 46 *American Politics Research* 465; Gauchat, 'Politicization of Science in the Public Sphere: A Study of Public Trust in the United States, 1974 to 2010' (2012) 77 *American Sociological Review* 167.

23 Kahan, Jenkins-Smith and Braman, supra n 18.

24 Pluviano, Della Sala and Watt, 'The Effects of Source Expertise and Trustworthiness on Recollection: The Case of Vaccine Misinformation' (2020) *Cognitive Processing* 1; Guillory and Geraci, 'Correcting Erroneous Influences in Memory: The Role of Source Credibility' (2013) 2(4) *Journal of Applied Research in Memory and Cognition* 201.

25 Benegal and Scruggs, 'Correcting Misinformation about Climate Change: The Impact of Partisanship in an Experimental Setting' (2018) 148 *Climate Change* 61; Bastardi, Uhlmann and Ross, 'Wishful Thinking: Belief, Desire, and the Motivated Evaluation of Scientific Evidence' (2011) 22 *Psychological Science* 731.

3. NORMATIVE STANDARDS FOR THE REGULATION OF FAKE NEWS

There is no general prohibition on the regulation of information on grounds of falsity under international law. The suppression of false or misleading information has been practiced in many jurisdictions in the context of, for example, defamation and sedition.[26] Each state enjoys sovereign prerogative to control and regulate the content of information created, published or disseminated within its jurisdiction, unless it is subject to a specific treaty obligation binding upon it, such as the obligation to respect and ensure respect for freedom of expression under Article 19 of the International Covenant on Civil and Political Rights (ICCPR).[27] Caution must therefore be exercised not to overreach international legal restrictions in non-signatory countries, such as Malaysia, Myanmar and Singapore, where national authorities enjoy greater freedom (unless it is argued that relevant ICCPR standards are customary international law). These countries have signed the non-binding ASEAN Human Rights Declaration, which affirms the right to freedom of opinion and expression, but that is subject to limitations 'for the purpose of securing due recognition for the human rights and fundamental freedoms of others, and to meet the just requirements of national security, public order, public health, public safety, public morality, as well as the general welfare of the peoples in a democratic society.'[28]

The normative issue arising from the regulation of fake news is, at the fundamental level, its compatibility with freedom of expression. This is so in so far as it concerns people's ability to generate, disseminate or receive certain content of information, whether that is a perceived fact or opinion, via a particular medium of communication. Article 19(3) of the ICCPR recognises that restrictions may be imposed on any form of expression or means of its dissemination 'as are provided by law and are necessary: (a) for respect of the rights or reputations of others; (b) for the protection of national security or of public order (*ordre public*), or of public health or morals.' This continues to apply in the digital context so that, as the UN Human Rights Committee observes, 'restrictions on the operation of websites, blogs or any other Internet-based, electronic or other such information dissemination system, including systems to support such communication, such as Internet service providers or search engines, are only permissible to the extent that they are compatible with paragraph 3 [of Article 19].'[29]

Therefore, under the legal regime where freedom of expression and information is guaranteed, the falsity of information alone cannot be a legitimate ground for restrictions.[30] Freedom of expression is not limited to 'correct' information and extends to

26 See generally Singh, *Sedition in Liberal Democracies* (2019); Kenyon, *Defamation: Comparative Law and Practice* (2006); Mitchell, *The Making of the Modern Law of Defamation* (2005).

27 1966, 999 UNTS 171. See also Article 32 Arab Charter on Human Rights 2004, reprinted in (2005) 12 IHRR 893; Article 9(2) African Charter on Human and Peoples' Rights 1981, 1520 UNTS 217; Article 13 American Convention on Human Rights 1969, 1144 UNTS 123; Article 10 Convention for the Protection of Human Rights and Fundamental Freedoms 1950, 213 UNTS 221.

28 2012, paras 8 and 23.

29 Human Rights Committee, General Comment No 34: Freedoms of opinion and expression (art. 19), 12 September 2011 at para 43 [hereinafter GC34]. See generally O'Flaherty, 'Freedom of Expression: Article 19 of the International Covenant on Civil and Political Rights and the Human Rights Committee's General Comment No 34' (2012) 12 *Human Rights Law Review* 627.

30 GC34, supra n 29 at para 49; *Joint Declaration on Freedom of Expression and 'Fake News', Disinformation and Propaganda*, adopted by the United Nations Special Rapporteur on Freedom of Opinion and Expression,

information and ideas that 'may shock, offend and disturb' people.[31] Freedom of political speech has traditionally enjoyed a privileged position particularly in democratic countries, where freedom of expression is considered an essential foundation for a democratic society and for its progress even if the information or ideas are offensive, shocking or disturbing.[32] Consistent with this approach, the signatories to the EU Code of Practice pledge that they should not delete or prevent access to otherwise lawful content 'solely on the basis that they are thought to be "false".'[33]

However, this does not mean that no restriction on the creation and dissemination of fake news is justifiable. Rather, national authorities are required to ensure that restrictions are justifiable on legitimate grounds as provided in the relevant treaty instrument. Thus, in *Paraga v Croatia*, the Human Rights Committee observed that criminal proceedings instituted against dissemination of false information 'may, *in certain circumstances*, lead to restrictions that go beyond those permissible under Article 19, paragraph 3' (emphasis added) of the Covenant.[34] The African Commission on Human and People's Rights also recognises a potential justification for regulating fake news *a contrario* when it has declared that no one shall be subject to sanctions or harm 'for releasing information on wrongdoing or which discloses a serious threat to health, safety or the environment, or whose disclosure is in the public interest, in the honest belief that such information is *substantially true*' (emphasis added).[35]

The extent to which, and circumstances in which, restrictions on the creation and dissemination of fake news is justifiable depends, therefore, on the applicability and the construction of a particular provision in which freedom of expression is guaranteed and as clarified in the subsequent development of jurisprudence. Although freedom of expression is variably formulated and construed in different jurisdictions, normative considerations are generally divided into the following three requirements: (i) the principle of legality; (ii) necessity; and (iii) proportionality.

the Organization for Security and Co-operation in Europe Representative on Freedom of the Media, the Organization of American States Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights Special Rapporteur on Freedom of Expression and Access to Information, 3 March 2017 at para 2(a).

31 *Joint Declaration*, supra n 30 at preamble para 7.

32 See, for example, *von Hannover v Germany (No 2)* Application Nos 40660/08, 60641/08, Merits and Just Satisfaction 7 February 2012 at para 101; *Handyside v United Kingdom* Application No 5493/72, Merits 7 December 1976 at para 49; C-274/99 *Connolly v Commission* [2001] ECR I-1638 at para 39; *Khushboo v Kanniammal* [2010] 5 SCC 600 at para 29 (Supreme Court of India); *Romesh Thappar v State of Madras* [1950] SCR 594 at 602 (Supreme Court of India); *Texas v Johnson*, 491 U.S. 397 at 414 (1989); *Rankin v McPherson*, 483 U.S. 378 at 387 (1987).

33 Section II.D *EU Code of Practice on Disinformation*, 26 September 2018, available at: ec.europa.eu/digital-single-market/en/news/code-practice-disinformation [last accessed 2 December 2020]. This Code of Practice was signed on a voluntary basis by representatives of online platforms, leading social networks, advertisers, and the advertising industry to address the spread of online disinformation and fake news. Signatories include Facebook, Google, Microsoft, Mozilla and Twitter.

34 *Paraga v Croatia* 727/1996, Views, CCPR/C/63/D/727/1996 (2001) at para 9.6.

35 Principle 35(1) *Declaration on Principles on Freedom of Expression and Access to Information in Africa* 2019, adopted by the African Commission on Human and Peoples' Rights.

A. The Principle of Legality

The first normative consideration is the procedural requirement that any restriction must be provided by law. Under the ICCPR, a restriction may be prescribed in various forms,[36] but must be formulated with sufficient precision to enable both individuals to regulate their conduct accordingly and those charged with its implementation to act in accordance with the law.[37] This means that any measure of interference with the creation and dissemination of fake news must be sufficiently clear regarding the content of information that will be subject to restriction.

However, there are difficulties in defining what constitutes fake news. The Protection from Online Falsehoods and Manipulation Act 2019 of Singapore regards a statement as false 'if it is false or misleading, whether wholly or in part, and whether on its own or in the context in which it appears.'[38] Although later repealed due to the change of government, Malaysia similarly enacted the Anti-Fake News Act in 2018, defining fake news as 'any news, information, data and reports, which is or are wholly or partially false' in any form capable of suggesting words or ideas.[39] Although the ICCPR does not apply in these jurisdictions, such broad definitions of fake news would not be amenable to objective assessment without a clear standard against which the statement can be verified as true or false.

Definitional difficulties with fake news have raised concerns about over-regulation and significant discretion that national authorities are granted due to an overly broad and vague manner in which restrictive measures are imposed. The UN Special Rapporteur, for example, criticises the vague terminology used in China's Cybersecurity Law,[40] for being 'so general as to permit officials excessive discretion to determine their meaning.'[41] According to the Special Rapporteur, the use of broad terms for criminal sanctions or the lack of specific conditions that justify blocking online content risks curtailing freedom of expression arbitrarily and excessively.[42]

Concerns about over-regulation are not unique to fake news. Indeed, similar concerns have been raised with Germany's Network Enforcement Act, which has imposed legal obligation on social media platforms to remove 'illegal content' such as defamatory statements.[43] Designed to combat hate speech, rather than fake news generally, the legislative measure with financial penalties for non-compliance has prompted social media platforms to err on the side of caution, blocking more content than is necessary.[44]

36 Cf GC34, supra n 29 at para 24.

37 Ibid. at para 25.

38 Section 2(2) Protection from Online Falsehoods and Manipulation Act 2019 (Singapore) [hereinafter POFMA].

39 Section 2 Anti-Fake News Act 2018 (Malaysia).

40 It requires network operators and users to 'observe public order' and 'respect social morality', proscribing creation or dissemination of 'false information to disrupt the economic or social order': Article 12 Cybersecurity Law 2016 (PRC).

41 Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc A/71/373, 6 September 2016 at para 13.

42 La Rue, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc A/HRC/17/27, 16 May 2011 at para 31.

43 Section 3 Netzwerkdurchsetzungsgesetz [Network Enforcement Act] 2017 (Germany).

44 Thomasson, 'Germany Looks to Revise Social Media Law as Europe Watches', *Reuters*, 8 March 2018, available at: www.reuters.com/article/us-germany-hatespeech/germany-looks-to-revise-social-media-law-as-europe-watches-idUSKCN1GK1BN [last accessed 2 December 2020].

Indiscriminate blocking which interferes with lawful content or websites as a collateral effect of the measure aimed at illegal content or websites can be considered arbitrary and as failing to satisfy the requirement of foreseeability.[45]

Legislative terms are inevitably couched in general terms, to a greater or lesser extent, which are to be clarified in practice through interpretation and application. 'Whilst certainty is desirable,' as the European Court of Human Rights observed in *The Sunday Times v United Kingdom*, 'it may bring in its train excessive rigidity and the law must be able to keep pace with changing circumstances.'[46] As such, European human rights jurisprudence adopts a flexible approach to the notion of foreseeability, which 'depends to a considerable degree on the content of the text in issue, the field it is designed to cover and the number and status of those to whom it is addressed.'[47] Thus, the ambiguity of the terms employed is not necessarily a bar to the regulation of fake news on account of foreseeability.

In particular, the requirement of foreseeability cannot be held to the same standard when the measure of interference concerns national security. In *C.G. v Bulgaria*, the European Court of Human Rights observed that this requirement does not go so far as to compel states to enact legal provisions listing in detail all conduct subject to regulation, acknowledging that 'threats to national security may vary and may be unanticipated or difficult to define in advance.'[48] What needs to be put in place instead is some form of process in which the allegation of falsity and its harmful nature can be challenged or independently assessed.[49]

It is therefore more likely that the requirement of foreseeability is satisfied when the regulation of fake news is qualified by requisite elements such as malicious intent and harm in a way that enables individuals to regulate their conduct accordingly and protect them against arbitrary interference. On the other hand, an unqualified fear or alarm would be too broad to satisfy this requirement.[50] The legislative amendment proposed in Taiwan reportedly defines 'disinformation' with three elements: being fake; motivated by malice; and harmful to individuals, organizations or social order.[51] And, as will be discussed below, various requisite elements such as knowledge, dishonesty, intention and likelihood of causing societal disturbances are commonly identified in defining fake news as an offence in multiple jurisdictions.

45 See, for example, *Vladimir Kharitonov v Russia* Application No 10795/14, Merits and Just Satisfaction, 23 June 2020 at paras 37–46; *Ahmet Yildirim v Turkey* Application No 3111/10, Merits and Just Satisfaction, 18 December 2012 at paras 57–68.

46 *The Sunday Times v United Kingdom* Application No 6538/74, Merits, 26 April 1979 at para 49. See also *Lindon, Otchakovsky-Laurens and July v France* Application No 21279/02, Merits and Just Satisfaction, 22 October 2007 at para 41; *VgT Verein Gegen Tierfabriken v Switzerland* Application No 24699/94, Merits and Just Satisfaction, 28 June 2001 at para 55; *Hertel v Switzerland* Application No 25181/94, Merits and Just Satisfaction, 25 August 1998 at para 35.

47 *Lindon, Otchakovsky-Laurens and July v France*, supra n 46 at para 41.

48 *C.G. v Bulgaria* Application No 1365/07, Merits and Just Satisfaction, 24 April 2008 at para 40.

49 *Malcolm Ross v Canada* 736/1997, Views, CCPR/C/70/D/736/1997 (2000) at para 11.4; *C.G. v Bulgaria*, supra n 48 at para 40.

50 *Chavunduka v Minister for Home Affairs* [2000] SC36 (Supreme Court of Zimbabwe) at 14.

51 Rickards, 'The Battle Against Disinformation', *Taiwan Business Topics*, 21 August 2019, available at: topics.amcham.com.tw/2019/08/battle-against-disinformation/ [last accessed 2 December 2020].

The UN Special Rapporteur appears to be concerned about the prospect of national authorities acting as 'arbiters of truth in the public and political domain,'[52] when vague terms such as false or misleading information allow them to exercise excessively broad discretion. However, the same Rapporteur appears to have greater confidence in the individual's ability to judge truth and assess what constitutes a threat or harm to the public interest, in the context of unauthorised disclosure of information as a whistle-blower.[53] On the other hand, Mark Zuckerberg, the founder of Facebook, calls for caution against online media service providers positioning themselves to be the arbiter of truth.[54] Given the diverse positions on the determination of falsity, it would be excessive to interpret this procedural requirement of legality to demand substantial clarity in how falsity in the content of information is to be established.

B. Necessity

Any restriction on freedom of expression must be necessary for the legitimate reasons specified in the relevant treaty, such as the protection of national security, public order, public health or morals. It is a substantive requirement involving two different aspects: the scope of restrictions (such as the content of information to be restricted); and the extent of restrictions (such as content removal or blocking standards and the severity of punishment in the case of conviction for a false or misleading statement).[55] The former concerns whether the legitimate interest can be protected in other ways that do not restrict freedom of expression,[56] whereas the latter addresses the test of proportionality as discussed in the next section.

The test of necessity requires the state parties to demonstrate in specific and individualised fashion the precise nature of the threat to the legitimate interest and the material link of the restrictive measure with the specific need for protection on which it is predicated.[57] This means that whether a particular measure of interference is necessary or not is assessed on the basis of different sets of circumstances in each case. The availability of and easy access to fake news in the digital age has serious repercussions for the ability of national authorities to maintain national security, public order, public health and morals. It is thus conceivable that restrictions on false or misleading information regarding health-threatening activities are deemed necessary so that, for example, the prohibition of false or misleading advertising of harmful substances is likely to be justifiable on public health grounds.[58]

52 See, for example, Kaye, *Letter from the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression to the Government of Italy*, 20 March 2018 at 4, available at: www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-ITA-1-2018.pdf [last accessed 2 December 2020].

53 Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc A/70/361, 8 September 2015 at para 63.

54 McCarthy, 'Zuckerberg Says Facebook Won't Be "Arbiters of Truth" After Trump Threat', *The Guardian*, 28 May 2020, available at: www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump [last accessed 2 December 2020].

55 Cf Nasu, 'State Secrets Law and National Security' (2015) 64 International & Comparative Law Quarterly 365 at 391.

56 See, for example, *Ballantyne, Davidson, McIntyre v Canada* 359/1989 and 385/1989, Views, CCPR/C/47/D/359/1989 and 385/1989/Rev.1 (1993) at para 11.4.

57 GC34, supra n 29 at paras 22 and 35.

58 See Taylor, *A Commentary on the International Covenant on Civil and Political Rights* (2020) at 574; Schabas, *Nowak's CCPR Commentary*, 3rd edn (2019) at 572.

The Human Rights Committee rejects the view that national authorities are granted a margin of appreciation in delimiting the scope of freedom of expression.[59] Rather, its jurisprudence suggests that national authorities must demonstrate the precise nature of the threat to a particular interest protected under the ICCPR in specific and individualised fashion.[60] For example, the Committee considered that the Republic of Korea failed to specify the precise nature of the threat allegedly posed by the expression of views sympathetic to the North Korean propaganda as the legitimate ground for criminal charges under the National Security Law.[61] The strict application of this approach, however, is likely to cause problems in justifying various measures to regulate fake news due to the diffused nature of the threats it is perceived to pose.

In the European context, guided by the principle of a democratic society and under the supervision of regional courts and other monitoring bodies, contracting states are granted a margin of appreciation in determining what is 'necessary' to protect the competing interests identified as the permissible grounds for restricting freedom of expression.[62] Thus, for example, criminal prosecution against Handyside, and the forfeiture and subsequent destruction of the book he published under the Obscene Publications Act 1959 (as amended in 1964), was found to be justifiable as a measure of interference that fell within the margin of appreciation, having regard to 'the different views prevailing [in each contracting state] about the demands of the protection of morals in a democratic society.'[63] When, on the other hand, the content of expression concerns a matter of public interest, the margin of appreciation is reduced.[64]

Generally, the margin of appreciation is considered narrow in relation to 'political expression,'[65] including expression on matters of public policy and concern, which requires a high level of protection under Article 10 of the European Convention on Human Rights. The Second Section of the European Court in *VgT* was therefore correct in reiterating that the margin of appreciation is reduced as far as the 'political

59 GC34, supra n 29 at para 36. Cf McGoldrick, 'A Defence of the Margin of Appreciation and an Argument for its Application by the Human Rights Committee' (2016) 65 *International & Comparative Law Quarterly* 21.

60 For a selection of cases in this regard, see Joseph and Castan, *The International Covenant on Civil and Political Rights: Cases, Materials, and Commentary*, 3rd edn (2013) at 613–17.

61 *Kim v Republic of Korea* 574/1994, Views, CCPR/C/56/D/574/1994 (1999) at paras 12.4–12.5.

62 See, eg, *Lindon, Otchakovsky-Laurens and July v France,* supra n 46 at para 45; *Hertel v Switzerland*, supra n 46 at para 46; *Vogt v Germany* Application No 17851/91, Merits, 26 September 1995 at para 52; *Handyside v United Kingdom*, supra n 32 at paras 48–49. For detailed analysis, see, for example, Bychawska-Siniarska, *Protecting the Right to Freedom of Expression under the European Convention on Human Rights* (Council of Europe 2017) at 44–62.

63 *Handyside v United Kingdom*, supra n 32 at para 57.

64 See, for example, *Bédat v Switzerland* Application No 56925/08, Merits, 29 March 2016 at para 49; *Morice v France* Application No 29369/10, Merits and Just Satisfaction, 23 April 2015 at para 125; *VgT Verein Gegen Tierfabriken v Switzerland*, supra n 46 at para 71; *Hertel v Switzerland*, supra n 46 at para 47. For various factors that affect the scope of the margin of appreciation, see Gerards, *The General Principles of the European Convention on Human Rights* (2019) at 172–96; Legg, *The Margin of Appreciation in International Human Rights Law: Deference and Proportionality* (2012) pt II.

65 See, for example, *Feldek v Slovakia* Application No 29032/95, Merits and Just Satisfaction, 12 July 2001 at para 74; *Sürek v Turkey (No 1)* Application No 26682/95, Merits and Just Satisfaction, 8 July 1999 at para 61; *Castells v Spain* Application No 11798/85, Merits and Just Satisfaction, 23 April 1992 at para 46; *Lingens v Austria* Application No 9815/82, Merits and Just Satisfaction, 8 July 1986 at para 42.

expression' is concerned.[66] On that basis the Court discounted the prohibition of political advertising that applied only to broadcasting media in Switzerland as being not 'of a particularly pressing nature.'[67] However, as the Court has recognised elsewhere, a somewhat wider margin of appreciation can be granted than normally afforded to restrictions in the public interest, where there is no European consensus on how to regulate an emerging or divisive problem.[68]

Indeed, the conventional standard that has applied to political expressions may be called into question in the diffused context of information disorder on social media. This is because there is no presumption of good faith on the part of content generators to provide accurate and reliable information in accordance with the tenets of responsible journalism.[69] Public concerns might prevail particularly when the content of expression has an effect of inciting violence against an individual, a public official or a sector of the population.[70] Assessment therefore requires an examination of the content of expression held against the person and the context in which it was expressed. This suggests that in Europe, the need for restrictions on freedom of expression varies depending on the topic that the fake news transpires; specifically, false statements that challenge or disturb political debates are more likely to be protected under Article 10 of the European Convention on Human Rights, in comparison to those that merely cause confusion and panic which threatens national security, public order, public health or morals.

C. Proportionality

Although not expressly provided, the test of proportionality has generally been considered to regulate the extent to which freedom of expression may be restricted on legitimate grounds.[71] Rather than justifying the need for regulation, it concerns the choice of means available to restrict the freedom, often requiring national authorities to choose the least intrusive measure of interference. In assessing whether the measure of interference is proportionate to the legitimate aims pursued, a number of different factors are taken into account: for example, the form of expression at issue as well

66 *VgT Verein Gegen Tierfabriken v Switzerland*, supra n 46 at para 71.

67 Ibid. at para 74.

68 *Animal Defenders International v United Kingdom* Application No 48876/08, Merits and Just Satisfaction, 22 April 2013 at para 123; *TV Vest AS and Rogaland Pensjonistparti v Norway* Application No 21132/05, Merits and Just Satisfaction, 11 December 2008 at para 67; *Wingrove v UK* Application No 17419/90, Merits and Just Satisfaction, 25 November 1996 at para 58.

69 *Times Newspapers Ltd v UK (Nos 1 and 2)* Application Nos 3002/03, 23676/03, Merits, 10 March 2009 at para 42; *Bédat v Switzerland*, supra n 64 at para 50; *Jersild v Denmark* Application No 15890/89, Merits and Just Satisfaction, 23 September 1994 at para 31. Cf *Steel and Morris v United Kingdom* Application No 68416/01, Merits and Just Satisfaction, 15 February 2005 at para 90 (extending safeguards more generally to those engaging in public debate); *Delfi AS v Estonia* Application No 64569/09, Merits and Just Satisfaction, 16 June 2015 at para 115 (extending the 'duties and responsibilities' to Internet news portals).

70 *Ceylan v Turkey* Application No 23556/94, Merits and Just Satisfaction, 8 July 1999 at para 34; *Sürek v Turkey (No 1)*, supra n 65 at para 61; *Chavunduka v Minister for Home Affairs*, supra n 50 at 18; *Brandenburg v Ohio*, 395 U.S. 444 at 447 (1969); *Schenck v United States*, 249 U.S. 47 at 52 (1919) (referring to 'clear and present danger').

71 GC34, supra n 29 at para 34.

as the means of its dissemination,[72] risk of abuse,[73] and the nature and severity of the sanctions imposed.[74] The test of proportionality must be met both in respect of the general measure of restriction (such as legislative measures requiring online media service providers to remove or block any false or misleading content) and the particular measure of interference with a specific content of expression.[75]

In the view of the European Court of Human Rights, the central question as regards restrictive measures of a general nature is not whether less restrictive measures should have been adopted, but rather whether, in striking the balance it did between competing rights and interests, the legislature acted within the margin of appreciation afforded to them.[76] The Court's approach to the prohibition of paid political advertising in broadcasting is particularly instructive in this respect. In *Animal Defenders International v United Kingdom*, the Court found it important that the prohibition was circumscribed to address the precise risk of distortion that paid advertising was capable of creating due to its inherently partial nature and the danger of unequal access based on wealth.[77] In the majority view, the general measure was specifically tailored to address the immediate, invasive and powerful impact of broadcasting media, of which the state was particularly wary.[78] However, views may be divided as to the precise boundary of proportionate restrictions,[79] which poses a particular challenge in prescribing the types of fake news that are subject to regulation.

Here, the distinction between statements of fact and value judgments drawn in the European human rights jurisprudence is instructive in identifying a proportionate boundary of limitation. In cases where a statement amounts to a value judgment, the European Court has suggested that 'the proportionality of an interference may depend on whether there exists a sufficient factual basis for the impugned statement.'[80] Such an approach indicates that the restriction of online content in the public interest is more likely to be justifiable when it is limited to the removal or blocking of content that lacks a sufficient factual basis and causes significant harm to the legitimate interests. On the other hand, the same measure is more likely to be subject to challenge when the statement made is purely a value judgment or the factual basis for the information is disputed.

72 Ibid. at para 34.

73 *Animal Defenders International v United Kingdom,* supra n 68 at para 108.

74 *Morice v France*, supra n 64 at para 127.

75 Jacobs, *The European Convention on Human Rights* (1975) at 201–2.

76 *Animal Defenders International v United Kingdom*, supra n 68 at para 110. See also *Evans v United Kingdom* Application No 6339/05, Merits, 10 April 2007 at para 91; *Mellacher v Austria* Application Nos 10522/83, 11011/84, 11070/84, Merits, 19 December 1989 at para 53.

77 Supra n 68 at para 117.

78 Ibid. at para 119.

79 See ibid. at Dissenting Opinion of Judge Tulkens, Joined by Judges Spielmann and Laffranque, para 12. For further analysis, see Lewis, 'Animal Defenders International v United Kingdom: Sensible Dialogue or a Bad Case of Strasbourg Jitters?' (2014) 77 *Modern Law Review* 460; Rowbottom, 'Animal Defenders International: Speech, Spending, and a Change of Direction in Strasbourg' (2013) 5 *Journal of Media Law* 1.

80 See, for example, *Morice v France*, supra n 64 at para 126; *Steel and Morris v United Kingdom*, supra n 69 at para 87; *Feldek v Slovakia*, supra n 65 at para 76; *Lingens v Austria*, supra n 65 at para 46.

4. EVALUATING REGULATORY RESPONSES

Traditionally, regulatory control of the content of information has taken the form of interlocutory injunctions and restraint orders preventing publication. With the invention of printing in the fifteenth century and its subsequent development to the arrival of mass media, various means of control have been exercised over certain means of expression at the dictates of the ruling authorities.[81] However, the development of the internet as the means of communication has outpaced the traditional methods of information regulation, with the exponential growth of user-generated content that can be easily shared and disseminated through social media.

As discussed above, the distributed, networked, and data-driven architecture of digital information technologies has amplified the psychological impact of social media for information consumption practices due to psychological biases (confirmation bias, and motivated cognition and information processing). This new medium of communication has turned the society into 'echo chambers'—where people primarily consume information from like-minded voices they have chosen—or 'filter bubbles'—where unseen algorithms select information likely to be preferred by the user. The existing regulatory framework that controls the sources of information is increasingly considered to be not fit for purpose in the digital society, where a vast amount of data is constantly generated and shared at any moment in time.[82]

The increased awareness that the dissemination of fake news is causing societal problems of information disorder has triggered the development of different regulatory responses with a view to controlling the content of information accessible to the public. The focus of regulation has shifted from the sources of information to the platforms where information is generated and disseminated, as well as internet users. The current regulatory responses to information disorder that are commonly observed at the national level can largely be grouped into three categories: information correction; content removal or blocking; and criminal sanction. For the purpose of this analysis, the complete shutdown of the internet is disregarded as a regulatory solution for it is clearly excessive and counter-productive to the preservation and restoration of orderly information system.[83]

A. Information Correction

The least intrusive form of regulation is information correction. This does not directly interfere with false or misleading information or access to it, but rather creates a designated digital platform where the falsity of a particular content is publicly announced. The idea is consistent with the ideology of the free flow of information in the 'marketplace of ideas' and is premised upon the assumption that individuals are rational enough to seek out truth when they come across a dubious content. Large social media platforms such as Twitter have also adopted information

81 See generally Siebert, *Freedom of the Press in England 1476–1776: The Rise and Decline of Government Control* (1952); Emerson, 'The Doctrine of Prior Restraint' (1955) 20 *Law and Contemporary Problems* 648 at 650–2.

82 *Guide to Guarantee Freedom of Expression*, supra n 14 at 21.

83 Milanovic, supra n 3 at Part III.

correction initiatives themselves, typically labelling content deemed to be 'synthetic' or 'manipulated.'[84]

In its March 2018 report, a High-Level Group of Experts, established by the European Commission to advise it on policy initiatives to counter false information on the internet, recommended various response measures that, among other things, enhance transparency of the digital ecosystem, promote information literacy, develop tools for empowering users and journalists, and calibrate the effectiveness of responses through continuous research.[85] Based on this recommendation, the European Union led the initiative of adopting the Code of Practice, in which the signatory companies commit themselves, inter alia, to invest in products, features and tools that facilitate content discovery and access to different sources representing diverse perspectives.[86]

The idea of establishing an online portal for correcting misinformation has been widely used in multiple countries, including Canada,[87] China,[88] Croatia,[89] Italy,[90] and Pakistan.[91] In the Democratic Republic of the Congo, where fake news was proliferating about the Ebola outbreak, the government has reportedly recruited young people to monitor and report misinformation circulating on WhatsApp—a major social media platform widely used in the country—for correction by communications experts with accurate information via WhatsApp or local radio.[92] Singapore has adopted a more interventionist approach with information correction. Under the Protection from Online Falsehoods and Manipulation Act 2019, Singaporean authorities can issue a correction direction to require a person who has made a false statement, or the internet intermediary service provider, to make a correction notice in the specified form and

84 See Twitter, 'Synthetic and Manipulated Media Policy' (undated), available at: help.twitter.com/en/rules-and-policies/manipulated-media [last accessed 2 December 2020].

85 European Commission, *Report of the Independent High Level Group on Fake News and Online Disinformation: A Multi-dimensional Approach to Disinformation* (2018) at 11 and 35–8, available at: ec.europa.eu/newsroom/dae/document.cfm?doc_id=50271 [last accessed 2 December 2020].

86 Section II.D *EU Code of Practice on Disinformation*, supra n 33.

87 Government of Canada, *Cabinet Directive on the Critical Election Incident Public Protocol*, 9 July 2019, available at: www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol/cabinet.html [last accessed 2 December 2020].

88 Qiu and Woo, 'China Launches Platforms to Stamp out "Online Rumors"', *Reuters*, 30 August 2018, available at: www.reuters.com/article/us-china-internet/china-launches-platform-to-stamp-out-online-rumors-idUSKCN1LF0HL [last accessed 2 December 2020].

89 'Croatian Government to Join EU Fight against Fake News', *Total Croatia News*, 21 January 2019, available at: www.total-croatia-news.com/politics/33682-fake-news [last accessed 2 December 2020].

90 *Italy's Remarks Following Communication from UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression on the Red Button Protocol, Mr. David Kaye*, May 2018 at 3, available at: www.ohchr.org/Documents/Issues/Opinion/Legislation/ItalyReplyMay2018.pdf [last accessed 12 August 2020].

91 'Govt Launches "Fake News Buster" Account to Expose False Reports', *Dawn*, 27 May 2020, available at: www.dawn.com/news/1436167 [last accessed 2 December 2020].

92 Spinney, 'Fighting Ebola Is Hard. In Congo, Fake News Makes It Harder', *Science*, 14 January 2019, available at: www.sciencemag.org/news/2019/01/fighting-ebola-hard-congo-fake-news-makes-it-harder [last accessed 2 December 2020].

manner.[93] This legislative scheme is designed to ensure that corrections are published alongside a false statement of fact so that the misinformed facts are corrected.[94]

From a normative perspective, information correction is relatively uncontroversial since, as noted above, it does not directly interfere with false or misleading information or access to it. However, the psychological mechanisms discussed above make it difficult to eliminate the adverse impact of fake news once it has been created, even where information is corrected. This is due to the susceptibility of individuals to: (i) believe fake news where it is consistent with pre-existing beliefs or values; and (ii) resist corrective information that contradicts them. These susceptibilities mean that once a person has been able to find fake news that is supportive of their existing beliefs or outlook it will be difficult to subsequently eliminate the influence of that fake news. Such resistance to information correction can be seen in response to influential misinformation, for example, the fake news linking autism to the MMR vaccine despite a systematic review by the US Institute of Medicine (now the National Academy of Medicine) finding no evidence for such a causal link.[95]

A large body of research, especially a 2017 meta-analysis of literature on information correction, compellingly demonstrates the unreliability of information correction in effectively eliminating misperceptions.[96] Results confirm the persistence of misinformation and suggest that people who generate arguments supporting misinformation are particularly likely to struggle later to question and change their initial attitudes and beliefs. Further studies examining the level of detail that information correction should contain to be most effective have shown mixed results, as additional detail seems to increase not only the effects of information correction but also misinformation persistence.[97] Although potential methods to increase the effectiveness of information correction have shown some promise in reducing the effects of misinformation, they have not been consistently validated and have not come close to eliminating the psychological effects and persistence of misinformation.[98]

Even worse, information correction may create a risk of what are known as 'backfire' effects, actually increasing people's commitment to endorsed misinformation. Research suggests that although such effects are rare, they may occur where content is particularly contentious, factual claims are ambiguous, or information correction strategies are not sufficiently robust.[99] Where backfire effects have occurred, this has been attributed to

93 Sections 11 and 21 POFMA 2019 (Singapore).

94 Neubronner, 'Bill to Protect Online Falsehoods: Refinements Needed', RSIS Commentary No 85, 30 April 2019, available at: www.rsis.edu.sg/rsis-publication/nssp/bill-to-protect-from-online-falsehoods-refinements-needed/#.Xs7CBW5FzD4 [last accessed 2 December 2020].

95 Poland, 'MMR Vaccine and Autism: Vaccine Nihilism and Postmodern Science' (2011) 86(9) *Mayo Clinic Proceedings* 869.

96 Man-pui et al, 'Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation' (2017) 28 *Psychological Science* 1531.

97 Ibid. at 1541–2.

98 See, for example, Kim, Moravec and Dennis, 'Combating Fake News on Social Media with Source Ratings: The Effects of User and Expert Reputation Ratings' (2019) 36 *Journal of Management Information Systems* 931.

99 Full Fact, 'The Backfire Effect. Does It Exist? And Does It Matter?' (2019), available at: fullfact.org/media/uploads/backfire_report_fullfact.pdf [last accessed 2 December 2020]. For examples of backfire effects, see Nyhan and Reifler, 'When Corrections Fail: The Persistence of Political Misperceptions' (2010) 32 *Political Behavior* 303; Hart and Nisbet, 'Boomerang Effects in Science Communication' (2012) 39 *Communication*

vigorous attempts to counter information contained in the correction, generating more attitudinally congruent information.[100] This observation is consistent with research suggesting that when a person generates explanations in line with misinformation, belief in that information is more persistent.[101]

The inefficacy of information correction strategies may be partially addressed through a shift to pre-emptive correction, such as psychological inoculation against misinformation or public communication efforts prior to its assimilation, which has shown some promise.[102] However, it should be noted that promising strategies reduce rather than eliminate belief in fake news and have not been tested and validated across contexts. Thus, while information correction will have a role in combating misinformation its effectiveness should not be exaggerated and, as the authors of a 2017 meta-analysis concluded, should be approached with low expectations.[103] This is particularly so in politically contentious areas where (i) adverse effects are most likely and (ii) the risk of backfire effects may be highest. Although normatively compliant, information correction is not reliably effective and can even be harmful due to potential 'backfire' effects.

B. Content Removal or Blocking

A more intrusive form of regulation to deal with fake news involves removing or blocking fake news content. The online content may be subject to removal or blocking by a team of content moderators who manually review contents that internet users have flagged as inappropriate or by means of automated content filtering.[104] Content removal or blocking is analogous to the traditional means of censorship and restraint orders, but can be tailored to target specific content. Concerns are therefore raised regarding this method of regulatory intervention when it is employed in an overly broad,

Research 701; Lewandowsky et al, 'Misinformation and Its Correction: Continued Influence and Successful Debiasing' (2012) 13(3) *Psychological Science in the Public Interest* 106; Cook, Ecker and Lewandowsky, 'Misinformation and How to Correct It' (2015) *Emerging Trends in the Social and Behavioral Sciences* 1.

100 Nyhan and Reifler, supra n 99 at 308.

101 Man-pui et al, supra n 96 at 1541.

102 See, for example, Pennycook et al, 'Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy Nudge Intervention' (2020) 31 *Psychological Science* 770; Roozenbeek and van der Linden, 'Fake News Game Confers Psychological Resistance Against Online Misinformation' (2019) 5 *Palgrave Communications* 1; Roozenbeek and van der Linden, 'The Fake News Game: Actively Inoculating Against the Risk of Misinformation (2018) 22 *Journal of Risk Research* 570; Select Committee on Communications, 'Growing up with the Internet' (2017) paras 76–85, available at: publications.parliament.uk/pa/ld201617/ldselect/ldcomuni/130/13007.htm#_idTextAnchor031 [last accessed 2 December 2020]; Cook, Lewandowsky and Ecker, 'Neutralizing Misinformation Through Inoculation: Exposing Misleading. Argumentation Techniques Reduces Their Influence' (2017) 12(5) *PLoS One* e0175799; van der Linden et al, 'Inoculating the Public Against Misinformation About Climate Change' (2017) 1(2) *Global Challenges* 1600008.

103 Man-pui et al, supra n 96 at 1544.

104 See Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, UN Doc A/HRC/38/35, 6 April 2018 at paras 32–38. For illustration of technological solutions that are explored in Europe, see Ignatidou, 'The Promise and Limitations of Technological Solutions to Disinformation', European Science-Media Hub, 20 March 2019, available at: sciencemediahub.eu/2019/03/20/the-promise-and-limitations-of-technological-solutions-to-disinformation/ [last accessed 2 December 2020].

vague or indiscriminate manner, on the basis of confidential blocking lists or without any possibility for review.[105]

Content removal or blocking may take the form of self-regulation at the initiative of online media service providers without legislative intervention. For example, Twitter has introduced a 'synthetic and manipulated media' policy, whereby 'tweets' containing 'deceptively altered media . . . in ways that mislead or deceive people about the media's authenticity where threats to physical safety or other serious harm may result' are subject to removal.[106] Also, in the context of the coronavirus crisis, both Twitter and Facebook have announced plans to remove potentially harmful false information.[107] Such policies mark a departure from the more general approach that these platforms have adopted, which focuses on labelling and downgrading rather than content removal.[108]

In several jurisdictions, a governmental agency has been established or designated with the power to remove or block online content. India's Information Technology Act 2000, for example, has allowed the central government to block access in the interest of sovereignty and integrity of India, defence of India, security of the State, friendly relations with foreign States or public order, and also for preventing incitement to the commission of any related offence.[109] In Indonesia, several government agencies are authorised to restrict online content under the Information and Electronic Transactions Law, provided that the restriction is in the public interest and intended to maintain public order.[110] As part of wide-ranging media regulation, Egypt introduced the amendments to the Media and Press Law in 2018, which has granted the Supreme Media Council the authority to suspend any website, blog or social media account that posts fake news.[111] In December 2018, France enacted legislation on combating the manipulation of information, which has authorised the Superior Audiovisual Council to suspend the broadcasting of service providers during the time of election campaign, or revoke their broadcasting rights, when they are found to be deliberately disseminating false information of a nature that compromises the fairness of the election, or prejudices the fundamental interest of the nation, under the influence of a foreign state.[112]

In other jurisdictions, intermediary companies are under an obligation to ensure that information available on their networks complies with national law and to remove or block online content containing false information.[113] In the European Union, on

105 See *Joint Declaration*, supra n 30 at para 3; Council of Europe Commissioner for Human Rights, 'The Rule of Law on the Internet and in the Wider Digital World', Council of Europe Issue Paper 2014 at 12–14, available at: rm.coe.int/ref/CommDH/IssuePaper(2014)1 [last accessed 2 December 2020]; La Rue, supra n 42 at para 31.

106 See Twitter, supra n 84.

107 'Coronavirus: Twitter Bans "Unsafe" Advice about the Outbreak', *BBC News*, 19 March 2020, available at: www.bbc.co.uk/news/technology-51961619 [last accessed 2 December 2020]; 'Facebook Is Removing Fake Coronavirus News "Quickly" COO Sheryl Sandberg Says', *CBS News*, 18 March 2020, available at: www.cbsnews.com/news/facebook-coronavirus-fake-news-coo-sheryl-sandberg/ [last accessed 2 December 2020].

108 See above text accompanying n 84.

109 Section 69A Information Technology Act 2000 (India).

110 Article 40 Law No 11/2008 (Indonesia).

111 Article 19 Law No 92/2016 as amended by Law No 180/2018 (Egypt).

112 Articles 33–1-1 and 42–6 Law No 86–1067 as amended by Law No 2018–1202 (France).

113 See, for example, Article 2 and 7(b) Order No 170 Inter-ministerial Prakas on Publication Controls of Website and Social Media Processing via Internet in the Kingdom of Cambodia (Cambodia); Articles

the other hand, intermediary companies are exempted from civil liability for unlawful content stored at the request of a recipient of the service unless they play an active role of such a kind as to give them knowledge of or control over the data stored.[114] In the USA, online media service providers are protected from civil liability under Section 230(c) of the Communications Decency Act for removing or moderating content that is deemed obscene or offensive, as long as this is done in good faith.[115] However, President Trump signed an Executive Order on 28 May 2020 to narrow the scope of this protection,[116] mounting pressure on the online media service providers to restrain themselves from removing or restricting access to online content that is not objectionable under this clause.

With the paradigmatic shift in pervasiveness and potency from broadcasting media to social media, it is conceivable that a general legislative measure authorising the removal or blocking of fake news is considered necessary for preventing the distortion of crucial public debates concerning matters involving national security, public order, public health or morals. Vietnam, for example, justifies various regulation under Decree 72 of 2013 by characterising the internet as 'an open environment, allowing users to search and provide information in such a free manner that, without laws and regulations applied, it can easily be abused to undermine traditional culture, moral, public safety and national security.'[117] Difficulties, however, arise under the ICCPR when national authorities attempt to demonstrate the precise nature of the threat to a particular interest protected in specific and individualised fashion.

In the European context, a margin of appreciation may be accorded to national authorities as they, in particular liberal democracies, struggle in finding the best solution to the problems that fake news is causing through social media to the integrity of democracy. It is, as Lord Bingham noted, 'reasonable to expect that democratically-elected politicians will be particularly sensitive to the measures necessary to safeguard the integrity of democracy.'[118] Indeed, the legislature is best placed to assess what general restrictions are necessary to ensure that the political process is not distorted by the dissemination of fake news through social media. However, the legislative authorization

15–16 Administrative Measures on Internet Information Services 2000 (PRC); Articles 9, 12(2) and 47 Cybersecurity Law 2016 (PRC); Sections 16, 28, 33, 43 POFMA 2019 (Singapore); Articles 5 and 25(6) Decree No 72/2013/ND-CP on the Management, Provision, Use of Internet Services and Online Information 2013 (Vietnam); Articles 5(1)(i) and 16(6) Law No 24/2018 Cybersecurity Law 2018 (Vietnam).

114 Article 14 Directive 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internet Market ('Directive on electronic commerce') [2000] OJ L 178/1; C-236/08 to C-238/08 *Google France SARL and Google Inc v Louis Vuitton Malletier SA and Others* [2010] ECR I-2417 at paras 114 and 120. For recent developments, see Gregorio, 'Expressions on Platforms: Freedom of Expression and ISP Liability in the European Digital Single Market' (2018) 3 *European Competition and Regulatory Law Review* 203.

115 47 US Code 230(c).

116 Section 2 Executive Order on Preventing Online Censorship 2020 (US).

117 *Letter dated 10 January 2014 from the Government of Vietnam to the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association, and Special Rapporteur on the Situation of Human Rights Defenders*, Ref 04/VNM.2014, 10 January 2014 at 2, available at: spcommreports.ohchr.org/TMResultsBase/DownLoadFile?gId=31983 [last accessed 2 December 2020].

118 *R (on the Application of Animal Defenders International) v Secretary of State for Culture, Media and Sport* [2008] UKHL 15 at para 33.

or endorsement of online content removal and blocking must be specifically tailored to address the distorting impact of fake news in a manner proportionate to the legitimate aim pursued. As discussed above, there are practical challenges in determining the precise boundary of proportionate restrictions, which depends on the topic that the fake news transpires and the extent to which factual accuracy can be verified.

In terms of effectiveness, removal of fake news is likely to suffer from some of the same problems as information correction strategies, particularly as the great speed at which fake news often spreads means that it is likely to have been viewed by a significant number of people prior to removal.[119] People who have been exposed to fake news that they have endorsed are unlikely to change their beliefs just because that information has been removed. Literature on mental models suggests that removal will be even less effective in changing beliefs than information correction, since in order to update beliefs people need a new way to understand information that replaces those beliefs.[120]

There is also a clear and demonstrated potential for adverse effects following the removal of information, most obviously through what is referred to as a 'Streisand effect'—a social phenomenon whereby the removal of content can actually draw increased attention to it.[121] Where people are committed to certain information, the removal of that information can be seen as proof of conspiracy or suppression of truth.[122] Research has suggested that in such cases removal can potentially speed up the spread of misinformation.[123] There is even a risk that this effect of removal will confer validity to similar misinformation remaining online.[124] This has the potential to be particularly problematic where content blocked in one jurisdiction is accessible in neighbouring countries, or available via virtual private network (VPN).

Blocking content prior to it ever appearing online is likely to be more effective, since it would avoid exposure of members of the public to misinformation. However, even if problematic content could reliably be identified before being posted, there would still be a significant risk of 'Streisand' effects where content consistent with beliefs or outlook is blocked. Such effects can be clearly seen in the US context, where the use of content moderation has been criticised by ideological conservatives who have claimed that content moderators are biased and censoring their viewpoints.[125] Here, blocking of information has a clear potential to promote conspiracy theories and, through Streisand effects, to reinforce beliefs that censored content represents the truth. As with content removal, this could be particularly problematic where content is accessible elsewhere.

119 See Vosiughi, Roy and Aral, 'The Spread of True and False News Online' (2018) 359 *Science* 1146.

120 Johnson-Laird, 'Mental Models and Probabilistic Thinking' (1994) 50 *Cognition* 180; Johnson and Seifert, 'Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences' (1994) 20 *Journal of Experimental Psychology* 1420; Wilkes and Leatherbarrow, 'Editing Episodic Memory Following the Identification Error' (1988) 40 *Quarterly Journal of Experimental Psychology* 361.

121 Jansen and Martin, 'The Streisand Effect and Censorship Backfire' (2016) 9 *International Journal of Communication* 656.

122 Ibid.

123 Del Vicario et al, 'The Spreading of Misinformation Online' (2016) 113(3) *Proceedings of the National Academy of Sciences* 554; Vosoughi, Roy and Aral, supra n 119.

124 Ting Chua et al, 'Identifying Unintended Harms of Cybersecurity Countermeasures' (2019) *APWG Symposium on Electronic Crime Research* 1 at 4.

125 Jiang, Robertson and Wilson, 'Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation' (2019) 13(1) *Proceedings of the International AAAI Conference on Web and Social Media* 278.

The demonstrated inefficacy of information correction cannot therefore be rectified by the removal or blocking of fake news. There are risks of inefficacy and backfire effects particularly where members of the public have been exposed to fake news either prior to removal or outside the national online environment. Due to the diffused nature of threats that fake news is perceived to pose, there are also challenges to normative justification where the need for removing or blocking content on grounds of falsity is strictly construed or where the factual basis for the information is disputed. These normative demands also pose practical challenges to intermediary online media service providers in complying with any legal requirement to remove or block fake news unless they engage in the editorial conduct to ensure that no statement of fact is generated without verification of its sources.

C. Criminal Sanction

Prosecution of fake news was established at the common law in as early as 1275 and continued as a subset of public mischief in the UK until the passage of the Public Order Act 1936.[126] In Cameroon, severe sanctions have been provided for publishing or reproducing any false statement, which is likely to bring public authorities into 'hatred, contempt or ridicule.'[127] Cote d'Ivoire has also criminalised '[l]a publication, la diffusion, la divulgation ou la reproduction par quelque moyen que ce soit, de Nouvelles fausses, de pieces fabruquées, falsifies ou mensongèrement attribuées à des tiers.'[128] There is a range of sanctions that may be imposed, from warnings for minor cases to re-education, disciplinary measures, fines and imprisonment for more serious violations.

The legislative move has recently geared towards criminalising the creation and dissemination of fake news in many jurisdictions.[129] Although later repealed, Malaysia enacted the Anti-Fake News Act in 2018, imposing criminal penalties for any person 'who, by any means, knowingly creates, offers, publishes, prints, distributes, circulates or disseminates any fake news or publication containing fake news.'[130] In Cambodia, an inter-ministerial directive ('Prakas') was issued to impose criminal penalty on

126 *De Scandalis Magnatum* 1275 3 Edw 1 at ch 34 ('from henceforth none be so hardy to tell or publish any false News or Tales, whereby discord, or occasion of discord or slander may grow between the King and his People, or the Great Men of the Realm'); UK Law Commission, *Working Paper No 84: Criminal Libel* (1982) at 10–13; *R v Zundel* [1992] 2 SCR 731 at 793–7 (Cory and Iacobucci JJ).

127 Law No 66/LF/13 (Cameroon); Sections 113 and 240 Panel Code, Law No 65/LF/24 (Book I) and Law No 67/LF/1 (Book II) (Cameroon). See also Fombad, 'Freedom of Expression in the Cameroonian Democratic Transition' (1995) 33 *Journal of Modern African Studies* 211 at 214–19.

128 Article 173 Law No 81–640 as amended by Law No 95–522 (Cote d'Ivoire).

129 See, for example, Section 25 Digital Security Act 2018 (Bangladesh); Article 30–2-3 Law on Mass Media 2008 as amended in 2018 (Belarus); Article 312–13 Penal Code 1996 as amended by Law No 25/2018 (Burkina Faso); Article 32 Criminal Law as amended on 29 August 2015 (PRC); Article 173 Law No 81–640 as amended by Law No 95–522 (Cote d'Ivoire); Article 80(d) Penal Code Law No 58/1937 as amended by Law No 95/2003 (Egypt); Sections 22–3 Computer Misuse and Cybercrimes Act 2018 (Kenya); Article 19 Decree No 327/2014 (Lao PDR); Section 68(a) Telecommunications Law 2013 (Myanmar); Sections 46(g) and 46(ga) Information and Communication Technologies Act 2001 as amended by Act No 21/2016 (Mauritius); Article 154(1) Act No 3815 Revised Penal Code 1930 (Philippines); Section 339–4 Criminal Code 1935 as amended on 18 June 2014 (Republic of China); Section 7 POFMA 2019 (Singapore); Section 361 Criminal Code 2005 (Slovakia); Section 14 Computer Crime Act (No 2) 2017 (Thailand).

130 Section 4(1) Anti-Fake News Act 2018 (Malaysia).

disseminating fake news.[131] In the USA, the State of Texas has made it a criminal offence, specifically a Class A misdemeanor, to create a 'deep fake video' (defined as a video created with the intent to deceive, which appears to depict a real person performing an action that did not occur in reality) and cause that video to be published or distributed within 30 days of an election.[132] California has also passed a law to combat the use of 'deep fakes' during election campaigns, but imposes civil rather than criminal liability for breach.[133]

These criminal sanctions are a promising regulatory solution, since they can target the initial creation and sharing of fake news, meaning that the public are never exposed to associated misinformation and disinformation. If fake news is not created in the first place, the risk of endorsement of and commitment to that news is eliminated. There is no specific empirical study demonstrating that the threat of criminal sanctions deters and eliminates the creation or dissemination of fake news. Rather, its purported impact is based on the conventional wisdom of criminal law rationalising the crime prevention effect of the threat of punishment.[134] This is what regulatory strategies utilising criminal sanctions, such as new Texas laws in relation to deep fakes during election periods, attempt to demonstrate.

It must be acknowledged that criminal sanctions would not be completely effective in eliminating fake news. This is because: (i) the threat of criminal sanctions does not deter everyone (especially those who are operating from foreign jurisdictions where it is not an offence) from criminal action; and (ii) this type of crime is relatively difficult to police due to technical problems common to law enforcement against various cybercrimes.[135] In addition, although this has not been empirically investigated, overbroad and undefined criminalization has a potential to produce Streisand effects, especially where a particular group feels that 'news' representing their viewpoints is specifically targeted and identified as false. Despite these potential limits, criminal sanction allows national authorities to mobilise law enforcement capacity to implement pre-emptive regulation, targeting attempts to create and distribute fake news.

The problem is rather their compatibility with normative requirements. The use of criminal sanctions for creating or disseminating false or misleading information has been considered disproportionate in several jurisdictions due to broad wording and the potential for arbitrary applications.[136] In *R v Zundel*, for example, the Supreme Court of Canada found that Section 181 of the Criminal Code infringed freedom of

131 Order No 170 Inter-ministerial Prakas on Publication Controls of Website and Social Media Processing via Internet in the Kingdom of Cambodia (Cambodia).

132 Texas Election Code s255.004 (effective from 1 September 2019).

133 California Election Code a20010 (effective from 3 October 2019).

134 See, for example, Paternoster, 'How Much Do We Really Know about Criminal Deterrence?' (2010) 100 *Journal of Criminal Law & Criminology* 765; Kennedy, *Deterrence and Crime Prevention, Reconsidering the Prospect of Sanction* (2009) at 9–10.

135 See, for example, Europol and Eurojust, 'Common Challenges in Combating Cybercrime', June 2019, available at: www.europol.europa.eu/publications-documents/common-challenges-in-combating-cybercrime [last accessed 2 December 2020]; Gottschalk, *Policing Cyber Crime* (2010) at 8; Collier and Spaul, 'Problems in Policing Computer Crime' (1992) 2 *Policing and Society* 307.

136 This, of course, depends on the different constitutional tests adopted in each jurisdiction. False news clauses have been upheld in *Public Prosecutor v Pung Chew Choon* [1994] 1 MLJ 566 at 578 (Edgar Joseph Jr SCJ), cited with approval in *Public Prosecutor v Azmi Bin Sharom* [2015] 6 MLJ 751 at para 37 (Federal Court of Malaysia).

expression,[137] although views were divided as to whether the legislative wording was too broad and more invasive than necessary to achieve the legitimate aim.[138] Likewise, the Supreme Court of Zimbabwe found an equivalent clause unconstitutional, observing that 'there are other means of achieving the impugned provision's aim far less arbitrary, unfair and invasive to free expression.'[139] More recently, the Supreme Court of India has struck down Section 66(A) of the Information Technology Act as unconstitutional for 'arbitrarily, excessively and disproportionately' invading freedom of expression, due to the broad wording enabling application for purposes not sanctioned by the Constitution.[140]

It cannot be ruled out, however, that a legislative provision criminalising the creation or dissemination of fake news can be tailored in a way that specifically, and with sufficient precision, addresses the legitimate interests despite the inherent ambiguity of falsity as the constitutive element of the offence. Indeed, the criminalization of fake news tends to target those who fabricate information for online dissemination and is often qualified by requisite elements such as knowledge,[141] dishonesty,[142] and the intention or likelihood to cause societal disturbances.[143] Rather than dismissing such provisions on account of ambiguity, the test of proportionality demands incorporating various safeguards, such as the requirement of malicious intent and the threshold of harm,[144] to restrict the extent to which criminal prosecution is used to protect the legitimate interests. Thus, for example, it is reasonable to consider necessary and proportionate criminally punishing someone who intentionally or knowingly creates and spreads fake news that incites violence or causes a public health crisis. The test of proportionality requires a commensurate level of threat to the legitimate interest that justifies this level of interference.

137 It reads: 'Everyone who wilfully publishes a statement, tale or news that he knows is false and causes or is likely to cause injury or mischief to a public interest is guilty of an indictable offence and liable to imprisonment . . . '.

138 Supra n 126 at 774 (McLachlin J), 824–42 (Cory and Iacobucci JJ).

139 *Chavunduka v Minister for Home Affairs*, supra n 50 at 22.

140 *Shreya Singhal v Union of India* [2015] SC 1523 at paras 82, 90 and 95.

141 Section 23 Computer Misuse and Cybercrimes Act 2018 (Kenya); Section 4(1) Anti-Fake News Act 2018 (Malaysia); Section 46(g) Information and Communication Technologies Act 2001 as amended by Act No 21/2016 (Mauritius); Section 59 Criminal Code Act 1990 (Nigeria); Section 7(1) POFMA 2019 (Singapore).

142 Section 68(a) Telecommunications Law 2013 (Myanmar); Section 14(1) Computer Crime Act (No 2) 2017 (Thailand).

143 Article 32 Criminal Law as amended on 29 August 2015 (PRC); Article 80(d) Penal Code 1937 as amended by Law No 95/2003 (Egypt); Section 23 Computer Misuse and Cybercrimes Act 2018 (Kenya); Section 46(ga) Information and Communication Technologies Act 2001 as amended by Act No 21/2016 (Mauritius); Section 59 Criminal Code Act 1990 (Nigeria); Article 154 Act No 3815 Revised Penal Code 1930 (Philippines); Section 32 Public Order Act 1965 (Sierra Leone); Section 7(1)(b) POFMA 2019 (Singapore); Section 63(a) Penal Code 1996 (Solomon Islands); Section 361 Criminal Code 2005 (Slovakia); Section 14 Computer Crime Act (No 2) 2017 (Thailand).

144 See Milanovic, supra n 3 at Part III. For similar considerations in the context of hate speech, see Kuhn, 'Reforming the Approach to Racial and Religious Hate Speech under Art 10 of the European Convention on Human Rights' (2019) 19 *Human Rights Law Review* 119 at 134–6. For a case study on civil and criminal regulation of Facebook and social media in the UK, see McGoldrick, 'The Limits of Freedom of Expression on Facebook and Social Networking Sites: A UK Perspective' (2013) 13 *Human Rights Law Review* 125.

Criminalization is more likely to meet the test of proportionality if it targets the 'methodology' behind the creation of the news, rather than an existing story involving value statements and accounts that people can become psychologically attached to. To that end, in contentious cases where falsity cannot be easily established without an iterative process of factual examination, people who are prosecuted as a result of publicising or disseminating online contents that are alleged to be false must have an opportunity to absolve themselves of liability by establishing that they are true or not known to be false. The Supreme Court of Zimbabwe, for example, adopted such methodological focus in *Chavunduka v Minister for Home Affairs*, in which it observed that '[f]ailure by the person accused to show, on a balance of probabilities, that any or reasonable measures to verify the accuracy of the publication were taken, suffices to incur liability even if the statement, rumour or report that was published was simply inaccurate.'[145] As has been established in the context of defamation,[146] the defence of truth and good faith would be critical as a safeguard against the abusive assertion of falsity.

Such a tailored approach also limits the risk of Streisand effects by setting criteria for methods and checks that must be employed by authors when creating and publishing an article in a news format. Due to the psychological biases discussed above, members of the public are less likely to object to requiring objective facts underlying news than they are to object to the blocking of a story they agree with. This means that an opportunity to contest allegations of falsity by presenting the factual basis for expressing a particular view is not only useful as part of various safeguards against abuse to meet the test of proportionality, but also necessary to ensure that criminal sanctions operate effectively to avoid increased attention to the suppressed information. Thus, the application of criminal sanctions in such a tailored manner has the potential to satisfy both normative and empirical criteria as a regulatory approach to combating fake news.

5. CONCLUSION

Developing regulatory solutions to combat fake news is challenging. Regulation must effectively combat the adverse effects of fake news while also respecting freedom of expression. This involves consideration of both normative principles and empirical realities, as well as the interaction between the two. The analysis in this article has shown that regulation cannot be reliably effective in eliminating the adverse effects of fake news without placing some level of restriction on freedom of expression. Therefore, an appropriate balance between combating fake news and respecting freedom of expression must be reached. This balance is likely to vary in different legal, political and cultural contexts, with divergence in normative constraint regarding the requirements of legality, necessity, and proportionality in different jurisdictions.

Information correction is the least intrusive form of regulation against fake news and does not infringe individual freedom of expression or access to information.

145 Supra n 50 at 15.

146 See, for example, *Morice v France*, supra n 64 at para 155; *Hasan Yazici v Turkey* Application No 40877/07, Merits and Just Satisfaction, 15 April 2014 at para 54; *Andrushko v Russia* Application No 4260/04, Merits and Just Satisfaction, 14 October 2010 at para 53; *Memère v France* Application No 12697/03, Merits, 7 November 2006 at paras 21–23; *Colombani v France* Application No 51279/99, Merits and Just Satisfaction, 25 June 2002 at para 66; *Castells v Spain,* supra n 65 at para 48.

In cases where fake news does not pose any harm to the legitimate interests, national authorities may have no choice but to rely upon information correction due to normative constraints they are subject to under the relevant rules of international or domestic law. Experts in psychology and communication are working to maximise the effectiveness of information correction.[147] However, psychological biases have been shown to make people resistant to information correction, particularly where fake news is consistent with their existing beliefs or cultural outlook. Current evidence on information correction suggests that it is not reliably effective and may have harmful 'backfire' effects.

The demonstrated inefficacy of information correction suggests that where stakes are high a more intrusive form of regulation needs to be explored. Removal and blocking of fake news is one potential approach. However, there are risks of inefficacy and backfire effects particularly where members of the public have been exposed to fake news either prior to removal or outside the national online environment. Due to the diffused nature of threats that fake news is perceived to pose, there are also challenges to normative justification where the need for removing or blocking content on grounds of falsity is strictly construed or where the factual basis for the information is disputed.

This means that the only way to effectively eliminate the effects of fake news would be to prevent the creation and distribution of such news in the first place. This article has identified criminal sanctions as an effective regulatory response due to their deterrent effect, based on the conventional wisdom of criminal law, against the creation and distribution of such news in the first place. Although further empirical research is needed, our analysis has demonstrated at least that criminalization of fake news in a context-specific fashion should not be dismissed on account of broad normative claims to worship freedom of expression. Indeed, criminal sanctions can be justified when the legal basis for imposing them is tailored in a way that specifically, and with sufficient precision, addresses the legitimate interests of national security, public order, public health or morals. Thus, for example, it is reasonable to consider necessary and proportionate criminally punishing someone who intentionally or knowingly creates and spreads fake news that incites violence or causes a public health crisis. The test of proportionality requires a commensurate level of threat to the legitimate interest that justifies this level of interference.

Both normative and empirical considerations should lead regulators to target a methodology for the creation of fake news lacking sufficient factual basis, rather than an existing story involving value statements and accounts that people can become psychologically attached to. Such a tailored approach helps avoid Streisand effects, since, for the reasons discussed above, people are unlikely to oppose a law requiring research and proper accounts underlying news in general, but would oppose the removal or blocking of a story giving an account they agree with (for example, linking the MMR vaccine and autism) even if it lacks sufficient evidence. This means that an opportunity to contest allegations of falsity by presenting the factual basis for expressing a particular view is not only useful as part of various safeguards against abuse to meet the test of proportionality, but also necessary to ensure that criminal sanctions operate effectively to avoid increased attention to the suppressed information.

147 See references cited supra nn 98 and 102.

Recognising that criminal regulation may be an effective way to target fake news does not mean that it should be used widely. Criminalising the creation and distribution of fake news means imposing a relatively severe sanction on a certain type of free expression. This involves social cost due to the chilling effect on the socially beneficial free flow of information. The extent to which social benefits of free flow of information are perceived to be outweighed by the public interest in the removal of the societal harms generated by fake news may vary, depending on how much social cost each society is prepared to accept. In authoritarian regimes, there would be greater risks of abuse to suppress dissent even if criminal sanctions are proven to be normatively justifiable and empirically effective.

The arguments presented in this article also have implications for intermediary liability; specifically, liability of companies that allow fake news to be posted on their online media platforms. Calling for action and screening by intermediaries is an important strategy currently gaining traction in multiple jurisdictions.[148] As certain types of false or misleading information become unlawful, the potential for intermediary liability is also likely to broaden. However, as discussed above, content removal and blocking is not likely to be reliably effective and could even be counter-productive, especially where members of the public have already been exposed to the ideas contained in the suppressed information.

The claims by non-governmental organizations notwithstanding,[149] intermediary liability for user-generated content has been established in multiple jurisdictions, where online media service providers are required to remove unlawful content.[150] The potential for intermediary liability can be limited by liability exemptions for providers publishing information generated or disseminated by third-party users.[151] The expansion of intermediary liability, on the other hand, is likely to generate incentives for online media service providers to censor a greater amount of content for efficient identification of fake news.[152] However, such liability cannot be justified unless the removal or blocking is limited to online content that lacks a sufficient factual basis and causes significant harm to the legitimate interests. This normative demand is likely to force intermediary companies to engage in editorial conduct to ensure that no statement of fact is generated without verification of its sources. Such demand will result in transforming online social media platforms into publishing companies susceptible to accusations of censorship and bias,[153] and may thus cause more harm to freedom of expression than the narrow criminalization discussed.

148 For analysis, see, for example, Spano, 'Intermediary Liability for Online User Comments under the European Convention on Human Rights' (2017) 17 *Human Rights Law Review* 665; Brunner, 'The Liability of an Online Intermediary for Third Party Content: The Watchdog Becomes the Monitor: Intermediary Liability after Delfi v Estonia' (2016) 16 *Human Rights Law Review* 163.

149 *Manila Principles on Intermediary Liability*, 24 March 2015, available at: www.eff.org/files/2015/10/31/manila_principles_1.0.pdf [last accessed 2 December 2020].

150 See Kaye, *Report of the Special Rapporteur on the Protection and Promotion of the Right to Freedom of Opinion and Expression: Overview of Submission Received in Preparation of the Report of the Special Rapporteur (A/HRC/38/35)*, UN Doc A/HRC/38/35/Add.1, 6 June 2018 at paras 7–11.

151 47 US Code 230(c).

152 *Guide to Guarantee Freedom of Expression*, supra n 14 at 22; *Joint Declaration*, supra n 30 at para 2(d).

153 See US Executive Order, supra n 116.

ACKNOWLEDGEMENTS

Preparation of this article was supported by a UK Research and Innovation Future Leaders Fellowship awarded to the first author. We are grateful to Professor Marko Milanovic, Dr Dimitrios Kagiaros, Dr Andrea Wallace, Dr Leanne Smith, and the anonymous reviewer for their comments on an earlier draft of this article.