International law and the challenge of disinformation: a patchwork of rights and obligations

Smulders, A.M.

5 Discriminatory Disinformation

5.1 INTRODUCTION

Throughout history, discriminatory speech and propaganda have served as precursors and catalysts for grave human rights violations and international crimes. The Armenian genocide, the Holocaust, the genocide in Rwanda and former Yugoslavia all demonstrate how false information, fabricated narratives and hateful rhetoric facilitate escalation from prejudice to violence.[1] The harmful consequences of persistent dehumanisation, vilification and degradation of groups and individuals affected by speech and enabled by the media, marked the 20th century. This pattern continues into the 21st century, though with increased sophistication in manipulated information and communication technology. The persecution of the Rohingya in Myanmar and the genocide of the Yezidi minority by the self-declared Islamic State (IS) exemplify this. Discriminatory speech scapegoating migrants, dehumanising and vilifying ethnic communities, and structurally stigmatising vulnerable gender minorities below the threshold of international crimes, is even more widespread.

Such discriminatory speech, generally referenced as 'hate speech', 'hate propaganda' or 'discriminatory propaganda', is the most versatile form of harmful speech subject to international law.[2] While universally condemned and unequivocally acknowledged as 'illegal content',[3] the parameters of permissible restrictions on such hate and atrocity speech remain ambiguous. Both international human rights law and international criminal law struggle with definitional and legal opaqueness, making the legal implications of the

1 Gregory Gordon, *Atrocity Speech Law* (Oxford University Press 2017).

2 In a human rights context, 'hate speech' prevails, while 'atrocity speech' or 'hate propaganda' generally refers to speech in the context of international crimes, in Markus P Beham, 'Atrocity Labeling' in Pavel Šturma and Milan Lipovský (eds), *The Crime of Genocide: Then and Now* (Brill Nijhoff 2022) 49; William Schabas, *Unimaginable Atrocities. Justice, Politics, and Rights at the War Crimes Tribunals* (Oxford University Press 2012).

3 European Commission, 'Strengthened Code of Practice on Disinformation' (16 June 2022) 1; UNGA, 'Disinformation and Freedom of Opinion and Expression During Armed Conflicts: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Kahn' (12 August 2022) UN Doc A/77/288, para 12; UNGA 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan' (7 August 2023) UN Doc A/78/288, para 24.

increasingly recognised synergy between these prohibited forms of speech and disinformation uncertain.[4]

This chapter explores these implications. It examines the convergence between disinformation and discrimination – captured by the notion of 'discriminatory disinformation' (5.2) – and to what extent international law governs the complex relationship between discriminatory speech, behaviour, and disinformation. Building on this foundation, section 5.4 examines the extent to which discriminatory disinformation falls within the scope of prohibited hate speech regulation under international human rights law, focussing on four categories of regulated speech: incitement to discrimination, incitement to hostility, dissemination of ideas of racial superiority and incitement to violence. Section 5.5 analyses the relationship between discriminatory disinformation and the prohibition of incitement to commit genocide under international criminal law. Limiting the analysis in the context of international crimes to genocide stems from the recognition that incitement to genocide is currently the only recognised substantive 'speech' crime.[5] This section positions discriminatory disinformation as a form of incitement but also explores whether disinformation containing denial of international crimes or false allegations of their occurrence, is subject to international regulation. This dimension challenges the boundaries of permissible restrictions on protected speech. Overall, the chapter advances the argument that while existing legal frameworks are imperfect, they provide grounds for regulating this phenomenon, conditional on nuanced application of established legal principles and adaptation to contemporary technological developments.

5.2 Discrimination, Disinformation and Violence

The relation between disinformation and discriminatory beliefs and behaviour is complex and multifaceted. From a cognitive and behavioural perspective, exposure to disinformation plays a significant role in shaping and facilitating the formation of discriminatory beliefs and ideas that precede legally prohibited behaviour such as incitement to violence. From a conceptual standpoint, when these discriminatory beliefs and ideas manifest as discriminatory speech, a convergence emerges between such speech and disinformation. This creates regulatory challenges that purely doctrinal approaches cannot adequately address. International legal frameworks currently lack well-defined boundaries and definitions for key concepts, including 'incitement' and 'hostility'.

4 UNSC, 'Resolution 2686 on Tolerance and International Peace' (14 June 2023) UN Doc S/RES/2686, preamble and para 10; UNGA, 'Disinformation and Freedom of Opinion and Expression During Armed Conflicts' (n 3) para 12.

5 Discriminatory disinformation as a mode of liability under international criminal law is discussed in section 6.3.1 'Modes of Liability'.

The concept of 'discriminatory disinformation' encapsulates the convergence between discriminatory speech and disinformation. It addresses a gap in legal taxonomy where these phenomena are traditionally treated separately despite their operational overlap. It enables more nuanced legal responses that capture speech producing discriminatory effects without explicitly targeting protected characteristics, creating a framework adaptable to evolving digital tactics. This term encompasses 'identity disinformation' – false or misleading information targeting individuals or groups based on their identity – as well as broader forms of disinformation that, while not explicitly identity-focused, nevertheless produce prohibited discriminatory effects (5.2.1).

Section 5.2.2 introduces the cognitive and behavioural mechanisms relevant to discriminatory disinformation. This focused analysis demonstrates how such disinformation instils and amplifies discriminatory beliefs, subsequently exacerbates these into discriminatory behaviour and potentially violence. Understanding these targeted mechanisms serves as a critical benchmark in delineating the processes of incitement to discrimination, hostility, and violence. Hence, this contextualisation is indispensable in identifying the problem and applying the international legal frameworks to discriminatory disinformation (5.3-5.5).

5.2.1 Discriminatory Disinformation

From the outset, discriminatory disinformation encapsulates any form of false or misleading information that is created, produced or disseminated with the intent to induce, forge or contribute to discrimination. This concept synthesises elements of targeted propaganda, false information, and discriminatory intent. While it partly coincides with the scope of 'hate speech' (5.2.1.1), 'discriminatory disinformation' is a separate, distinct phenomenon, which can be identified through the same methodological approach as in chapter four (4.3). The 'discriminatory' component of such disinformation can be identified with reference to the object of the message (5.2.1.2) and/or the objective of the actor who creates, produces or disseminates it (5.2.1.3). Moreover, discriminatory disinformation is more than the convergence of disinformation and prohibited discriminatory speech. Discrimination serves as a catalyst for disinformation – as existing patterns of discrimination serve as fertile ground for the creation and proliferation of disinformation – and disinformation is frequently instrumentalised to spread and sustain discriminatory ideas and behaviour. This mutual reinforcement sustaining discriminatory disinformation is simultaneously enabled by and creates an information environment that is fundamentally distorted. Without this distortion, discriminatory disinformation would not be the problem it has become. (5.2.1.4).

5.2.1.1 Convergence of 'Hate Speech' and 'Disinformation'

Two paradigmatic obstacles prevail in addressing discriminatory speech, including disinformation. The first is its rich, but confusing vocabulary in international discourse; 'hate speech', 'incitement to hatred', 'hate propaganda', 'verbal assaults', 'linguistic violence' and 'assaultive speech' are used largely interchangeably to describe expressions that are hateful and/or that encourage violence towards a group or its members.[6] Second, though 'hate speech' has emerged as the most prevalent term,[7] this term is often artificially separated from disinformation in policy and legal frameworks.[8] This separation stems from positioning disinformation as a 'residual category' of harmful speech, applicable exclusively when expression falls outside already regulated forms of speech. However, as illustrated in figure 4 and argued below, this separation is artificial and hinders, rather than advances, regulating discriminatory speech all together.

6 Lynne Tirell, 'Genocide Language Games' in Ishani Maitra and Mary Kate McGowan (eds), *Speech & Harm: Controversies over Freech* (Oxford University Press 2012) 176; Katharine Gelber, 'Differentiating Hate Speech: A Systemic Discrimination Approach' (2021) 24 Critical Review of International Social and Political Philosophy 421, 396.

7 Tarlach McGonagle, *Minority Rights, Freedom of Expression and the Media: Dynamics and Dilemmas* (Intersentia 2011) 318; James Jacobs and Kimberley Potter, *Hate Crimes: Criminal Law and Identity Politics* (Oxford University Press 1998) 11.

8 European Commission, 'A Multi-Dimensional Approach to Disinformation – Report of the independent High level Group on fake news and online disinformation' (March 2018) Directorate-General for Communication Networks, Content and Technology, 5 ('[d]isinformation] does not cover issues arising from the creation and dissemination online of illegal content (notably defamation, hate speech, incitement to violence'); Claire Wardle and Hossein Derakhshan, 'Information Disorder: Towards and in Disciplinary Framework for Research and Policy Making' (2017) Council of Europe report DGI(2017)09, 20; Claire Wardle, 'A Conceptual Analysis of the Overlaps and Differences between Hate Speech, Misinformation and Disinformation' (June 2024) United Nations Office of the Special Adviser on the Prevention of Genocide (OSPG) 4; UNGA, 'Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance, Ashwini KP' (18 October 2023) UN Doc A/78/538, para 16-17 (while recognizing 'there is a nexus between online racist hate speech on the one hand and disinformation and misinformation on the other, in particular in an online context', she 'asserts that disinformation and misinformation are distinct from online racist hate speech'); Peter BMJ Pijpers, 'On Cognitive Warfare: The Anatomy of Disinformation' (21 March 2024) The Defence Horizon Journal ('disinformation – in contrast to malinformation (hate speech […])'); UNSG, 'United Nations Strategy and Plan of Action of Hate Speech (May 2019) 15('the least severe form[s] of hate speech' and 'must not be subject to legal restrictions under international law' […] '*unless* such forms of expression also constitute incitement to hostility, discrimination or violence under article 20 (2) of the International Covenant on Civil and Political Rights [emphasis added]'); United Nations, 'The Guidance on Hate Speech Related to COVID-19' (11 May 2020) 2 (disinformation 'is closely linked' to hate speech, without providing any further clarification).

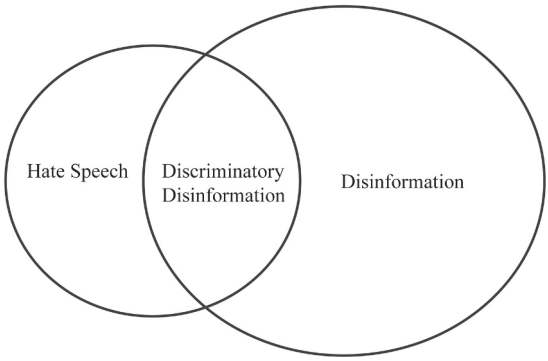

Figure 4: Discriminatory Disinformation and Hate Speech

The convergence between disinformation and unlawful hate speech principally determines the extent to which international law governs discriminatory disinformation. While neither 'hate speech' nor 'disinformation' is explicitly defined in international law, the two phenomena intersect in at least three key dimensions:

1. *Shared Rhetorical Strategies and Content*: identified instances of hate speech and disinformation employ identical patterns of dehumanisation, stigmatisation and threat construction. Contemporary analysis reveals that what is labelled 'hate speech' and 'disinformation' frequently exhibits similar language patterns and rhetorical strategies to target vulnerable groups.[9] Classifications maintaining a distinction contrastingly rely on inconsistent, asymmetric, framing: while hate speech is typically described by its specific content (dehumanisation, slurs, coded language) disinformation is defined by its cognitive effects (deception, undermining trust).[10]
2. *Common Harm Mechanism*: both hate speech and disinformation erode social coherence and threaten democratic stability – harms often (exclusively)

9 Mohsen Mosleh, Rocky Cole and David G Rand, 'Misinformation and Harmful Language are Interconnected, Rather than Distinct, Challenges' (2024) 3 PNAS Nexus 1-4, 1; Erik C Nisbet, 'The Psychology of State-Sponsored Disinformation Campaigns and Implications for Public Diplomacy' (2019) 14 The Hague journal of Diplomacy 65, 67-71.

10 Wardle (n 8) 9

associated with disinformation –[11] while disinformation simultaneously catalyses harms traditionally associated with hate speech.[12] A harm-based dichotomy ignores contemporary information ecosystems wherein disinformation's harm extends far beyond democracy and its derivatives to human rights, public health, climate change, science and beyond, equally manifesting at the individual level.[13]

3. *Complex Relationship to Truth:* the traditional position that hate speech contains '[i]nformation that is based on reality'[14] while disinformation is inherently false oversimplifies their relationship to truth.[15] Hate speech 'is most likely to occur when information is found to be completely false,'[16] and even when its informational foundation is purportedly truthful, in practice it consistently transgresses into misleading information through decontextualisation, exaggeration or disproportionate amplification of isolated events. Limiting hate speech to truthful information also creates a legal inconsistency. In applying hate speech regulation, it is widely recognised that 'no one should be penalized for statements that are true' – establishing a clear 'defence of truth'.[17] The very existence of this defence indicates that hate speech is not conceptually limited to, nor predominantly consists of, 'information based on reality.'

Contesting an artificial separation of disinformation and hate speech, however, does not equate to suggesting merging them under a single umbrella term of 'harmful speech'. While not identical, their differences do not logically necessitate mutually exclusive treatment under international law. Regulatory frameworks that treat these phenomena as entirely separate categories risk creating protection gaps that sophisticated information operations can exploit. Rather, this analysis advocates for an integrated approach that recognises their

11 Michal Bilewicz and Wiktor Soral, 'Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization' (2020) 41 Political Psychology 1, 3-33; Lilian Kojan *et al.*, 'Defend Your Enemy. A Qualitative Study on Defending Political Opponents Against Hate Speech Online' in Max van Duijn *et al.*, (eds), *Disinformation in Open Online Media* (Springer International Publishing 2020) 80-94.

12 Bilewicz and Soral (n 11) 3-33.

13 Section 1.5.1.4 'Likelihood of Harm'.

14 Wardle and Derakhshan (n 8) 5, 20; Pijpers (n 8).

15 Mal-information occurs when 'people are […] targeted because of their personal history or affiliations […] [w]hile the information can sometimes be based on reality (for example targeting someone based on their religion) the information is being used strategically to cause harm', in Wardle and Derakhshan (n 8) 20.

16 Michael Hameleers *et al.*, 'Civilized Truths, Hateful Lies? Incivility and Hate Speech in False Information – Evidence From Fact-Checked Statements in the US' (2020) 25 Information, Communication & Society 11, 1596-1613.

17 UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue' (7 September 2012) UN Doc A/67/357, para 50; Amal Clooney and David Neuberger (eds), *Freedom of Speech in International Law* (Oxford University Press 2024) 198.

harmful *modus operandi*, preventing fragmented and siloed legal and policy responses. Despite assertions that legal and historical contexts necessitate treating hate speech and disinformation as distinct phenomena,[18] their interconnected harms and complex relationship to truth call for frameworks acknowledging their fluid boundaries. Having this understanding, the following sections will examine the specific dimensions of discriminatory disinformation: the object of the message, the objective of the author, and the information environment in which it operates.

5.2.1.2 Object of the Message

Discriminatory disinformation encompasses false or misleading information targeting the identity or treatment of a particular group, intentionally created, produced or disseminated to cause harm to that group or its members. Building on Reddi's concept of 'identity propaganda', this category encompasses 'identity *disinformation*' as harmful content that weaponises falsehoods about groups' identity.[19] Identity disinformation operates through specific psychological and social mechanisms and consistently adapts to exploit social tensions across various contexts – from public health crises and migration debates to historical revisionism. It often consists of dehumanising narratives that demonise marginalised communities (as seen in anti-refugee propaganda linking migrants to disease or terrorism,[20] or gendered disinformation).[21] It

18 Wardle (n 8) 5, 7.

19 Madhavi Reddi *et al.*, 'Identity Propaganda: Racial Narratives and Disinformation' (2023) 25 New Media & Society 8, 2201-2218; European External Action Service, 'OSINT Guidelines: How to Detect and Analyse Identity-Based Disinformation/FIMI' (November 2024) 10.

20 Kimberly Grambo, 'Fake News and Racial, Ethnic, and Religious Minorities: A Precarious Quest for Truth' (2019) 21 University of Pennsylvania Journal of Constitutional Law 1299, 1300; Melanie Radue, 'Harmful Disinformation in Southeast Asia: "Negative Campaigning", "Information Operations" and "Racist Propaganda" – Three Forms of Manipulative Political Communication in Malaysia, Myanmar, and Thailand' (2019) 18 Journal of Contemporary Eastern Asie 2, 69.

21 UNGA 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Kahn' (n 3) para 12 (gendered disinformation comprises disinformation which 'targets women and gender nonconforming individuals, because of the gendered nature of the attacks and their gendered impact, and, very importantly, because it reinforces prejudices, bias and structural and systemic barriers that stand in the way of gender equality and gender justice'); Ric Neo and Jason Dean-Chen Yin, 'Of Social Discipline and Control' (2023) 31 International Journal on Minority and Group Rights, 110-132, 114-115; Nina Jankowitz *et al.*, 'Malign Activity, How Gender, Sex, and Lies are Weaponized Against Woman Online' (Wilson Centre Science and Technology Innovation Program, January 2021) 1-3; Judit Szakacs and Eva Bognar, 'The Impact of Disinformation Campaigns Abouts Migrant and Minority Groups in the EU' (European Parliament Directorate General for External Policies of the Union, June 2021) 12-21; European Commission, Communication From The Commission to the European Parliament and Council 'A More Inclusive and Protective Europe: Extending the List of EU Crimes to Hate Speech and Hate Crime' (9 December 2021) COM(2021) 777 final, para 3.4.2.

also manifests in seemingly 'positive' but equally harmful narratives and characterisations, such as myths about racial immunity to diseases that lead to healthcare discrimination and structural exclusion.[22]

Other narrative strategies include instrumentalising 'erasure strategies', denying the very existence or legitimacy of certain identities – exemplified by State policies refusing to recognise minority groups, as Australia did in relation to, *inter alia*, the Aboriginal peoples during the drafting process of Article27 ICCPR.[23] These patterns frequently involve historical revisionism and denial of systemic racism. Striking examples include false narratives linking refugees to disease spread or terrorist activities, anti-Asian sentiments and general 'ethnicisation' of crises during and following the COVID-19 pandemic,[24] and 'racial hoaxes' portraying minority groups – mostly migrants, Sinti and Roma – as economic or cultural threats.[25] At the extreme end stand disinformation narratives promoting ideologies such as "white genocide" or replacement conspiracy theories – political myths rooted in pseudoscience and pseudohistory that purvey the idea of a conspiracy to decimate white people globally to legitimise violence against non-white groups.[26] Rather than being spread as singular stories, these narratives appear simultaneously and systematically exploit identity-based vulnerabilities and rely on false or misleading information, collectively and gradually inducing discriminatory outcomes.

---

22 Will Mittendorf, 'Racist and Antiracist Conspiracy Theories' (2024) Inquiry: An Interdisciplinary Journal of Philosophy, 13-14; Jason G Randall, 'Improving Contact Tracing in Minority Communities by Combatting Misinformation and Distrust' (Understanding and Eliminating Minority Health Disparities In a 21st-Century Pandemic: A White Paper Collection 2021); Oluwadamilola Aiyeewumi and Malachy Ifeanyi Okeke, 'The Myth that Nigerians Are Immune to SARS-CoV-2 and that COVID-19 Is A Hoax Are Putting Lives At Risk' (2020) 10 Journal of Global Health 2; Janell Ross, 'Coronavirus outbreak revives dangerous race myths and pseudoscience' NBC News (19 March 2020); Reuters, 'False Claim: African Skin Resist the Coronavirus' Reuters(11 March 2020); Linda Villarosa, 'Myths About Physical Racial Differences Were Used to Justify Slavery – And are Still Believed by Doctors Today' The New York Times Magazine (14 August 2019); Michael Ruane, 'A Brief History of the Enduring Phony Science That Perpetuates White Supremacy' The Washington Post (30 April 2019).

23 Dieter Kugelmann, 'The Protection of Minorities and Indigenous Peoples Respecting Cultural Diversity' (2007) 11 Max Planck Yearbook of United Nations Law 233-263, 246.

24 Szakacs and Bognar (n 21) vi, 13-15, 18-20; Apoorvanad, 'How the Coronavirus Outbreak in India Was Blamed on Muslims' Al Jazeera (18 April 2020); Zamira Rahin, 'In the Latest Sign of Covid-19 Related Racism, Muslims Are Being Blamed for England's Coronavirus Outbreaks' CNN (6 August 2020); Human Rights Watch, 'COVID-19 Fuelling Anti-Asian Racism and Xenophobia Worldwide' (12 March 2020); Hannah Ellis-Petersen and Shaikj Azizur Rahman, 'Coronavirus Conspiracy Theories Targeting Muslims Spread in India' The Guardian (13 April 2020).

25 Gustavo Ferreira Santos, 'Misinformation and Hate Speech' in Oscar Pérez de la Fuente *et al.*, *Minorities, Free Speech and the Internet* (Routledge 2023) 123-124; Samira Abraham *et al.*, 'The Spatial Drivers of Discrimination: Evidence From Anti-Muslim Fake News in India' (Quaderni – Working Paper DSE No 1180,11 Januari 2023); BBC, 'EU Blasts Hungary 'Fake News' on Migrants' BBC (19 February 2019).

26 Section 5.5.4.2. '"White Genocide" Conspiracy Theories as Discriminatory Disinformation'.

5.2.1.3 Objective of the Author

Discriminatory disinformation also comprises disinformation of which the objective of the actor who creates, produces or disseminates the narratives is to discriminate or cause discriminatory consequences. Unlike other forms of disinformation, such as defamatory or terroristic disinformation, the intent behind discriminatory disinformation is often driven by emotion rather than being explicitly, or even exclusively, result oriented. In general, three primary objectives can be identified:

1. To influence personal sentiments creating discriminatory attitudes.
2. To escalate these sentiments into acceptance and motivation of low-level violence or discriminatory practices.
3. To directly incite severe or large-scale violence.

These objectives are not necessarily linear, and not all forms of discrimination aim to culminate in imminent violence. Many disinformation campaigns and operations serve as political tools, furthering anti-immigration agendas,[27] diverting attention from other contentious issues or influencing inter-State relations.[28] Some instances operate at an even higher level of abstraction, such as the spread of 'racialised disinformation': the intentional exploitation of wedge issues related to race, racial justice, or communities of colour to preserve existing political and social power structures.[29]

While this does not primarily engage with disinformation that is spread exclusively for profit – as explained in chapter one, the economic motivations behind discriminatory disinformation cannot be overlooked. Throughout history, racial and other forms of discriminatory disinformation have been instrumentalised to suppress and exploit vulnerable groups for commercial gain. From colonial-era slavery propaganda to modern practices of labour exploitation, a persistent pattern of racial disinformation serving as a smokescreen for economic exploitation emerges.[30] At a systemic level, the contemporary profit-driven nature of social and online media contributes to this prolifera-

27 Elein Culloty and Jane Suiter, 'Anti-Immigration Disinformation', in Howard Tumber and Silvio Waisbord (eds), *The Routledge Companion to Media Disinformation and Populism* (Routledge 2021) 221-230.

28 World Health Organization, 'WHO-Convened Global Study of Origins of SARS-CoV-2: China Part' (14 January – 10 February 2021); Yanzhong Huang, 'What the WHO Investigation Reveals About the Origins of COVID-19' Foreign Affairs (31 March 2021).

29 Technology and Social Change Project, 'Racialized Disinformation' (2019-2023) *The Media Manipulation Case Book*, Accessed 15 July 2024; Reddi (n 19) 2201-2218.

30 Michelle Amazeen *et al.*, 'Missing Voices: Examining How Misinformation-Susceptible Individuals From Underrepresented Communities Engage, Perceive, and Combat Science Misinformation' (2023) 46 Science Communication 1, 24; Shireen Mitchell, 'Disinformation: A Racist Tactic, from Slave Revolts to Election' (4 November 2022) Blog Union of Concerned Scientists, Accessed 22 August 2024; Africa Centre for Strategic Studies, 'Mapping a Surge of Disinformation in Africa' (Infographic,13 March 2024).

tion. As established previously,[31] the monetisation of the information system thrives on engagement and exposure, which are highest for polarising, sensationalist, and bias-exploiting content.[32] Consequently, discriminatory disinformation narratives often become the most visible and persistent,[33] contributing to a systematically discriminatory information environment.

5.2.1.4 Information Environment

The information environment serves both as context for and catalyst of discriminatory disinformation. By mirroring societal norms and attitudes, media and information ecosystems may amplify prevailing prejudices, inequality and systemic racism.[34] While some outlets and platforms have positively impacted the fight against discrimination, inherent biases and inequalities present in wider society are often reproduced within the information production and dissemination landscape.[35]

These structural inequalities in the information ecosystem manifest in different forms. First, there are significant disparities in internet and media accessibility between certain geographical areas, creating digital divides aligning with existing social and economic inequalities. This directly affects the availability of trustworthy and credible sources for marginalised and underrepresented communities.[36] Second, this environment creates asymmetric exposure to disinformation: vulnerable and marginalised groups are more

31 Section 1.3 'Societal Context of Disinformation'.

32 Sol Hart *et al.*, 'Politicization and Polarization in COVID-19 News Coverage' (2020) 52 Science Communication 5, 679-697; Amazeen *et al.* (n 30) 23, 27.

33 Grambo (n 20) 1317-1320.

34 Amazeen *et al.* (n 30) 7; Radue (n 20) 71; Reddi (n 19) 2202; Nolan Higdon and Alison Butler, '(Fake)News is Racist: Mapping Culturally Relevant Approaches to Critical News Literary Pedagogy' (2023) 14 Critical Education 3, 78-90; LaGarrett King, 'Don't Believe the Hype: Black History, the Media and Fake News' (2020) 2 The International Journal of Critical Media Literacy 149-173; Jessica Jaiswal *et al.*, 'Disinformation, Misinformation and Inequality-Driven Mistrust in the Time of COVID-19: Lessons Unlearned from AIDS Denialism' (2020) 24 AIDS and Behavior 2776-2780; Council of Europe, 'Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law' (5 June 2024) CETS No 225, preamble para 6 (reflecting that the parties are '[c]oncerned about the risks of discrimination in digital contexts, particularly those involving artificial intelligence systems, and their potential effect of creating or aggravating inequalities, including those experienced by women and individuals in vulnerable situations, regarding the enjoyment of their human rights and their full, equal and effective participation in economic, social, cultural and political affairs').

35 Rachel Kuo and Alice Marwick, 'Critical Disinformation Studies: History, Power and Politics' (2021) 2 Harvard Kennedy School of Misinformation Review 4.

36 Bilewicz and Soral (n 11) 3-33.

frequently targeted by disinformation.[37] As explained by Amazeen in the context of minorities in the United States, 'communities of colour are often specifically targeted by […] to obstruct their societal enfranchisement.'[38] This disparity is also strikingly illustrated by the Russian interference during the 2016 US Presidential elections, which revealed disproportionate targeting and impersonation of black communities.[39] During the COVID-19 pandemic communities of colour, particularly women, more likewise intensively exposed to medical mis- and disinformation.[40] This asymmetry extends to nearly all kinds of disinformation, political, health-related or otherwise.[41]

A 'disinformation environment' also amplifies the underlying distrust towards authorities and institutions already marginalised groups often have,[42] making them sceptical towards corrections from public authorities. Disinformation tactics, including impersonation and *sockpuppetry* online,[43] further discredit sources indispensable to countering disinformation causing their discrimination.[44] The combination of asymmetrical exposure, amplified scepticism and the prevailing socio-economic disparities in education, access to health care and political representation, results in minority groups and marginalised communities being systemically and disproportionately affected by this mutually reinforcing dynamic of disinformation and discrimination online.[45] These circumstances directly influence disinformation's legality, as further explained in section 5.4.4.3.

37 Amazeen *et al.* (n 30) 7; Higdon and Butler (n 34) 78-90; King (n 36)149-173; Ans Irfan *et al.*, 'Misinformation, Health Equity, News Media: Application of Critical Race Theory (CRT) to Examine News Media's Role in Normalizing Religious Bigotry' (2021) 26 Harvard Public Health Review.

38 Amazeen *et al.* (n 30) 7.

39 Szakacs and Bognar (n 21) 9; Deen Freelon *et al.*, 'Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation' (2022) 40 Social Science Computer Review 3, 560-578.

40 Randall (n 23) 1-44; Jaiswal *et al.*, (n 34) 2776-2780; Dhanaraj Thakur and Madrigal DeVan Hankerson, 'Facts and their Discontent: A Research Agenda for Online Disinformation, Race and Gender' (Center for Democracy and Technology, 11 February 2021); Samuel Woolley, 'In Many Democracies, Disinformation Targets the Most Vulnerable' (Centre for International Governance Innovation, 18 July 2022); Sheera Frenkel, 'Black and Hispanic Communities Grapple with Vaccine Misinformation' The New York Times (10 March 2021).

41 Amazeen *et al.* (n 30) 4; Jaiswal *et al.*, (n 34) 2776-2780.

42 *Ibid.* 4, 8.

43 *Sock puppetry* refers to the creation of a false online personae to conceal one's identity and motives, in Freelon *et al.* (n 39) 562.

44 UNGA, 'Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance' (n 8) para 20; UNGA, 'Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance: Racial Discrimination and Emerging Digital Technologies: A Human Rights Analysis, E Tendayi Achiume' (18 June 2020) UN Doc A/HRC/44/57, para 25.

45 Jaiswal *et al.*, (n 34) 2776-2780.

5.2.2 Cognitive and Behavioural Mechanisms of Discriminatory Disinformation

Among the various cognitive and behavioural processes exploited by disinformation – analysed in chapter one – two phases are particularly significant in the context of discriminatory disinformation: the exploitation of normal cognitive functions to instil and amplify discriminatory beliefs (5.2.2.1) and their subsequent amplification and escalation, leading to discriminatory behaviour and violence (5.2.2.2). These cognitive and behavioral mechanisms directly inform three key legal determinations. First, understanding how discriminatory disinformation operates helps identify when content reveals intent to advocate hatred or incite prohibited outcomes. Second, the cognitive impact of discriminatory disinformation informs whether expressions create a likelihood of discrimination, hostility, or violence – a central element in determining when speech restrictions are justified. Third, cognitive vulnerability factors help evaluate how disinformation as a contextual factor influences this likelihood, an equally relevant factor for these assessments.

5.2.2.1 Foundations of Discriminatory Disinformation

The formation of discriminatory beliefs stems from several interrelated cognitive processes on information reception and selection that disinformation strategically exploits.[46] Drawing from the broader insights from chapter one, these primary mechanisms are pressured by discriminatory disinformation. First, this contributes to cognitive overload in an information-saturated information environment. While in 1998, the UN Special Rapporteur of the Commission on Human Rights expressed his concern about the presence of '[o]ver 200 sites worldwide [that] are disseminating racist propaganda,'[47] this has increased by a factor of millions. As cognitive resources are depleted, individuals are more likely to rely on stereotypes and heuristics, increasing susceptibility to simplified discriminatory narratives, which facilitates and sustains discriminatory patterns.

Second, discriminatory disinformation creates states of confusion, where stereotypes are more likely to be applied. Stereotypes are 'cognitive shortcuts' that become particularly potent in ambiguous situations where information is lacking or contradicting narratives are presented.[48] By strategically introducing false information, propagating conflicting narratives, and distorting collect-

46 Moshe Hirsch, 'Cognitive Sociology, Social Cognition and Coping with Racial Discrimination in International Law' (2020) 30 European Journal of International Law 4, 1319-1338, 1325.

47 UNGA, 'Note by the Secretary General: Measures to Combat Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance' (17 August 1998) UN Doc A/53/269, para 29.

48 Hirsch (n 46) 1330.

ive perceptions of reality, disinformation campaigns cultivate an environment of uncertainty. This artificially created ambiguity then serves as fertile ground for the proliferation of stereotypes and, consequently, discriminatory attitudes,[49] which ultimately become self-reinforcing.

Third, it emphasises social distinctions and highlights inter-group differences. When disinformation targets a particular individual or group based on their identity, it employs reductionist narratives, distilling their identity into singular, often simplified stereotypical aspects. This oversimplification serves to portray the targeted individuals or groups as fundamentally 'other', distinct and separate from the perceived in-group.[50] This classification process transgresses from identifying them as 'others', to 'us versus them' and culminates in a moral judgement of 'good versus bad'. By consistently emphasising the division, the ability to empathise with 'the other' erodes.[51] This exploitation does not require active engagement with discriminatory disinformation; its harmful impacts can manifest even in passive recipients.[52] Social sciences have demonstrated that even mere exposure to discriminatory speech or content can have deleterious effects on, *inter alia*, inter-group empathy. This challenges traditional doctrinal understandings on the relation between speakers and audience in the legal framework on prohibited incitement.

5.2.2.2 From Cognition to Action

These cognitive insights and the stereotyping, polarising and inflammatory false narratives about minority groups characterise discriminatory disinformation and constitute the initial elements of incitement.[53] The most salient question for legal analysis, however, pertains to disinformation's role in facilitating the transition from rhetoric to action. The progression from disinformation-induced attitudes to violent action is a complex process that intersects individual psychology, group dynamics, and broader societal shifts.[54] Drawing from social science studies, three cognitive and behavioural patterns, or phases, emerge:[55]

1. Individual-level transformation: personal sentiments evolve from dislike through animosity to hate, building upon pre-existing cognitive biases.

---

49 Michael Hogg, 'Uncertainty, Group Identification and Intergroup Behavior: Positive and Negative Outcomes of how People Experience Uncertainty' (2024) PsyHub Special Issue, 63-72.

50 Wibke Timmermann, *Incitement in International Law* (Routledge 2015) 26.

51 *Ibid.*; Ervin Staub, *The Psychology of Good and Evil* (Cambridge University Press 2003) 120.

52 Agnieszka Pluta, 'Exposure to Hate Speech Deteriorates Neurocognitive Mechanisms of the Ability to Understand Others' Pain' (2021) 13 Scientific Reports 4127.

53 Timmerman (n 50) 29.

54 Richard A Wilson, *Incitement on Trial* (Cambridge University Press 2017) 229; Russell H Fazio, 'Multiple Processes by which Attitudes Guide Behavior: The Mode Model as an Integrative Framework' (1990) 23 Advances in Experimental Social Psychology 75-109.

55 Staub (n 51) 289, 303.

2. Group-level mobilisation: sentiments begin to motivate and justify discriminatory conduct and low-level violence, marking a shift from individual psychological change to broader group-level transformations.
3. Societal escalation: a behavioural turning point or period occurs when large scale violence emerges, coinciding with widespread communal or societal polarisation.[56]

While difficult to analyse in the abstract, loosely identifiable patterns of narrative and rhetorical features characterise this progression. Recalling that disinformation reinforces negative inter-group perceptions through stereotyping and stigmatisation, it subsequently contributes to shifting attitudes that perceive members of the 'other group' as legitimate targets of aggression.[57] For example, the fabrication or false attribution of violent incidents to minority groups significantly amplifies the risk of retaliatory actions and facilitates the psychological justification for violence.[58] The 'Lisa case' in Germany exemplified this dynamic: a fake news story spread by Russia about the rape of a Russian-German girl by Arab migrants fuelled the anti-immigration movement and led to large scale violent demonstrations.[59] Similarly, false accusations of engagement in ritualistic practices, such as child sacrifice, have targeted cultural or religious groups.[60] Fabricated religious prophecies claiming that, *inter alia*, Muslims will imminently wage a "holy war" against Christians have incited violence for centuries and constructions of 'Great Replacement' theories create perceived existential threats to justify violence.[61]

Disinformation's efficacy in these processes is magnified by its operational characteristics in contemporary information environments. Characterised by commercial incentives for sensationalist content, online and secluded communities, and a digital infrastructure of unprecedented scale, reach, and speed of content, discriminatory disinformation becomes highly pervasive.[62] Extensive and repetitive exposure to low-key inflammatory and discriminatory content

56 Anthony Oberschall, 'Propaganda, Hate Speech and Mass Killings' in Predrag Dojćinović (ed), *Propaganda, War Crimes Trials an International Law: From Speaker's Corner to War Crimes* (Routledge London 2012) 174; Antoine Buyse, 'Words of Violence: "Fear Speech," or How Violent Conflict Escalation Relates to Freedom of Expression' (2014) 36 Human Rights Quarterly 4, 779-797, 782.

57 Wilson (n 54) 34; Alexander Tsesis, *Destructive Message: How Hate Speech Paves the Way for Harmful Social Movements* (New York University Press 2002) 86.

58 Buyse (n 56) 785; Donald Horowitz, *The Deadly Ethnic Riot* (University of California Press 2001) 555.

59 Jakub Janda, 'The Lisa Case: STRATCOM Lessons from European States' (Security Policy Working Paper No. 11/2016, 2016); Welt, 'Russlanddeutsche demonstrieren gegen „Ausländergewalt' Welt (25 January 2016).

60 Jeroen Temperman, *Religious Hatred and International Law* (Cambridge University Press 2016) on religious hatred

61 Mittendorf (n 22) 10.

62 Section 1.3 'Societal Context of Disinformation'.

gradually reshapes beliefs, attitudes and subsequent behaviour, yet also creates risks of unexpected escalation, D'Alessandra and Gilea argue, where 'the chain of events leading to the onset of mass violence can be shortened, making atrocities increasingly unpredictable and prevention efforts more difficult.'[63] While the metaphorical effect of the 'hammer bashing on people's heads'[64] and 'spreading petrol little by little'[65] applies to disinformation generally, the evidence establishing a correlation between inciting discriminatory speech online and offline harms is 'overwhelming.'[66]

5.2.3 Interim Conclusion

'Discriminatory disinformation' is a complex, multifaceted phenomenon that operates at the intersection of falsehood and prejudice and emerges where disinformation and hate speech converge. Artificial taxonomic distinctions between disinformation and hate speech ignore this reality, which is fundamental to understanding how established legal frameworks prohibiting hate speech can and should be applied to discriminatory disinformation. Grasping discriminatory disinformation requires analyses of the object of the message (targeting group identity through negative stereotyping, manipulative "positive" stereotyping, or identity denial), the objective of the author (ranging from influencing attitudes to inciting violence) and the contextual information environment that enables its proliferation. Each of these dimensions reflects discriminatory disinformation's distinct and dangerous nature.

Discriminatory disinformation represents more than merely false information about minorities or discriminatory content – it constitutes a mechanism through which existing social inequalities are systematically reinforced and exploited. The mutually reinforcing relationship between discrimination and disinformation creates a dangerous cycle: pre-existing prejudices provide fertile

63 Federica D'Alesandra and Ross James Gildea, 'Technology, R2P, and the UN Framework of Analysis for Atrocity Prevention' (2024) 16 Global Responsibility to Protect 363, 379, 390.

64 Susan Benesch, 'Vile Crime or Inalienable Right: Defining Incitement to Genocide' (2008) 48 Virginia Journal of International Law 3, 524; *Prosecutor v. Tadic* (Case No. IT-94-I-|T) [Opinion and Judgement] (7 May 1997) para 96.

65 *Ibid*. 524; *Prosecutor v. Nahimana* (Case No. ICTR 99-52-T) [Judgement and Sentence] (3 December 2003) para 1099; Richard Delgado and Jean Stefancic, *Must We Defend Nazis?* (New York University Press 1997) 4-5.

66 Talita Dias, 'Finding Common Ground: The Right to Be Free from Incitement to Discrimination, Hostility, and Violence in the Digital Age' (2024) 16 Global Responsibility to Protect 391, 396; UNHRC, 'Report of the Independent Fact-Finding Mission on Myanmar' (2018) UN Doc A/HRC/39/64, para 73-74; Matteo Cinelli, *et al.*, 'Dynamic of Online Hate and Misinformation' (2021) Nature Scientific Reports, 11; Jason Chan, Anindya Ghose, and Robert Seamans, 'The Internet and Racial Hate Crimes: Offline Spillovers from Online Access' (2016) 40 MIS Quarterly 2, 381-403.

ground for disinformation, while disinformation simultaneously intensifies discrimination. This understanding establishes the foundation for meaningful legal analysis, highlighting gaps in traditional doctrines prohibiting discriminatory speech when applied to discriminatory disinformation. Similar to the other chapters, the transformed information environment and sophisticated cognitive exploitation stand at the core of the misalignment between 20th century regulation and 21st century sociotechnological phenomena.

5.3 Legal Continuum of Discriminatory Disinformation

The instrumentalisation of hate speech and disinformation far predates the 20th century,[67] yet prohibitions of discriminatory speech in international law emerged as a direct response to the Nazi propaganda that was central to the WWII atrocities. This regulatory framework developed along two parallel tracks: international human rights law and international criminal law. While the Nuremberg Trials established the precedent for the prohibition of discriminatory propaganda and disinformation in relation to incitement to genocide – later codified in the Genocide Convention and expanded by the International Criminal Tribunal for Rwanda and the International Criminal Tribunal for former Yugoslavia –[68] most discriminatory expressions fall below this threshold. This notwithstanding, they may still form a valid ground for restricting speech under international human rights law – particularly Articles 19 and 20 ICCPR, and Article 4 ICERD.

This analysis examines how these existing legal frameworks govern 'discriminatory disinformation', proceeding in two parts: the first examines how established prohibitions on incitement to violence, hostility and discrimination under international human rights law – referred to as 'unlawful hate speech'– apply to discriminatory disinformation (5.4). It engages with the International Covenant on Civil and Political Rights (ICCPR) and the Convention on the Elimination of All Forms of Racial Discrimination (ICERD). A third treaty regulating incitement, the Convention on the Prevention and Punishment of the Crime of Genocide, is included in section 5.5. This part analyses the applicability of the prohibition of incitement to commit genocide under international criminal law as well as the applicability of international criminal and human rights law to disinformation that exploits denial of genocide or fabricates false allegations of genocide.

67 Gordon (n 1) 31-36.

68 Eric De Brabandere, 'Propaganda' (2019) Max Planck Encyclopaedia of Public International Law, para 22.

5.4 Unlawful Hate Speech in International Law

While a uniform legal definition remains elusive, 'unlawful' hate speech – discriminatory expressions that are prohibited or subject to legitimate restriction under international law – consists of four specific categories:

1. Incitement to discrimination
2. Incitement to hostility
3. Dissemination of ideas of racial superiority
4. Incitement to violence

These prohibitions derive primarily from Article 20(2) ICCPR alongside Article 19(3), and Article 4 ICERD, while also reflected in Article 13(5) of the American Convention on Human Rights and the African Charter of Human and Peoples' Rights.[69] Discriminatory disinformation as a form of *unlawful* hate speech is thus regulated under international law respectively proscribing that certain expressions *must* be prohibited and/or *may* be restricted. This section engages with the first two, the International Covenant on Civil and Political Rights (ICCPR) and the Convention on the Elimination of All Forms of Racial Discrimination (ICERD).[70]

The first paragraph of Article 20 ICCPR declares that 'all propaganda for war shall be prohibited'. The second addresses hate speech, unequivocally stating that '[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.'[71] Article 4 of the ICERD is textually more elaborate, calling upon States in its preamble to:

> 'condemn all propaganda and all organizations which are based on ideas or theories of superiority of one race or group of persons of one colour or ethnic origin, or which attempt to justify or promote racial hatred and discrimination in any form, and undertake to adopt immediate and positive measures designed to eradicate all incitement to, or acts of, such discrimination and, to this end […].'

Sections (a) to (c) subsequently contain three obligations for States. First, to '[…] declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any

69 American Convention on Human Rights (adopted 22 November 1969, entered into force 18 July 1978) 1144 UNTS 123, Article 13(5); African Charter on Human and Peoples' Rights (adopted 27 June 1981, entered into force 21 October 1986) 1520 UNTS 217, Article 28.

70 UNGA, 'Universal Declaration of Human Rights' (adopted 10 December 1949) 217 A (III), Article 7 equally requires States to provide protection against incitement to discrimination.

71 International Covenant on Civil and Political Rights [hereafter: ICCPR] (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171, Article 20(2).

assistance to racist activities, including the financing thereof.' Second, to '[…] declare illegal and prohibit organizations, and also organized and all other propaganda activities, which promote and incite racial discrimination, and shall recognise participation in such organizations or activities as an offence punishable by law' and third, to 'not permit public authorities or public institutions, national or local, to promote or incite racial discrimination.'[72]

The ICERD is both narrower and broader in scope than the ICCPR. The Convention is narrower because it *prima facie* only covers expressions related to race, colour or ethnic origin. Moreover, where Article 20 ICCPR refers to 'incitement to discrimination, hostility or violence', Article 4 ICERD does not include 'hostility'. Article 4, however, covers a broader range of expressions that must be 'declared an offence punishable by law,' including '*all dissemination of ideas* based on racial superiority or hatred' as well as '*all* […] *propaganda activities*, which promote and incite racial discrimination.' (emphasis added).[73]

All measures implementing these obligations – criminal or otherwise – must meet the Article 19(3) ICCPR tripartite test of legality, legitimacy and necessity;[74] disproportionate or unnecessary measures result in a violation of Article 19(3). The appropriate regulatory response depends on the circumstances, a State's capacity to act as well as the provision: Article 4 ICERD requires criminal sanctions for prohibited expressions, while States enjoy greater discretion under Article 20(2) ICCPR, permitting criminal, civil or administrative penalties.[75] Both provisions are, however, subject to a clear international consensus that criminal penalties are justified only in exceptional circumstances and should be reserved for the most extreme cases of hate speech.[76]

Beyond prescribing when States *must* restrict hate speech, Article 19(3) ICCPR provides a framework for when States *may* do so.[77] Article 19(3) ICCPR in conjunction with Articles 2(1), 3 and 26 ICCPR allows for speech to be

72 International Convention on the Elimination of All Forms of Racial Discrimination [hereafter: ICERD] (adopted 21 December 1965, entered into force 4 January 1969) 660 UNTS 195, Article 4.

73 *Ibid.*

74 In Article 4 IICERD this is proscribed in the 'due regard' clause; Manfred Nowak, *U.N. Covenant on Civil and Political Rights – CCPR Commentary* (3rd rev edn, 2019 NP Engel) 583; *Mohamed Rabbae, ABS and NA v The Netherlands* Communication No 2124/2011 (29 March 2017) UN Doc CCPR/C/117/D/2124/2011 para 10.4.

75 Amal Clooney and Alice Gardoll, 'Hate Speech' in Amal Clooney and David Neuberger (eds), *Freedom of Speech in International Law* (Oxford University Press 2024) 175 ft. 170; Dias (n 66) 396.

76 *Ibid.* 204.

77 Arguably, this appears to be the preferred avenue for the Human Rights Committee, addressing most individual communications through the framework of legitimate restrictions under Article 19 rather than as direct violations of Article 20, in Mona Elbahtimy, *The Right to be Protected from Incitement to Hatred* (Cambridge University Press 2021) 97; Clooney and Gardoll (n 75) 183.

restricted if this is necessary 'for respect of the rights or reputations of others' or 'for the protection of national security or of public order (*ordre public*) or of public health or morals.'[78] As elaborated in chapter one,[79] these restrictions must be 'provided for law', constitute 'the least intrusive instrument among those which might achieve their protective function' and remain 'proportionate to the interest protected'.[80]

Sections 5.4.1-5.4.6 centralise prohibited expressions under Article 20(2) ICCPR and Article 4 ICERD. Beyond their literal text, these provisions contain substantial definitional and interpretative complexities. Article 20(2) ICCPR prohibits *advocacy* of *hatred* that *constitutes incitement* to three categories: discrimination, hostility and violence targeting race, nationality and/or religion. Similarly, Article 4 ICERD encompasses ambiguous forms of expression, such as the *promotion of ideas* related to racial superiority or hatred, incitement and 'all propaganda activities that promote racial discrimination'. The extent to which discriminatory disinformation falls within the ambit of these prohibitions hinges on the precise interpretation of these elements. Therefore, the chapter systematically analyses these provisions in relation to discriminatory disinformation by first identifying the protected groups under both articles (5.4.1), then examining 'advocacy' in Article 20(2) ICCPR, including its relationship with intent (5.4.2). Following, the notion of 'hatred' as articulated in both 'advocacy for hatred' (ICCPR) and 'dissemination of ideas [...] based on hatred' (ICERD) is clarified (5.4.3). This examination reveals that legal prohibition of hateful speech consistently depends on whether speech amounts to 'incitement' (5.4.4), either to non-physical harms (5.4.5) or violence (5.4.6).

5.4.1 Target Group

Discriminatory disinformation may constitute hate speech when targeting individuals or groups based on their identity characteristics. While identity encompasses numerous factors – ranging from ethnicity and sexual orientation to lifestyle and personal interests – international law narrowly circumscribes unlawful forms of hate speech to factors of nationality, race and religion (Article 20(2) ICCPR) or race, colour and ethnic origin (Article 4 ICERD), notably omitting other grounds such as sexual orientation, gender, age or political opinion.

78 ICCPR (n 71) Article 20(2).
79 Section 1.4.1 'Freedom of Expression'.
80 UN Human Rights Committee, 'General Comment No 34: Article 19, Freedoms of Opinion and Expression' (12 September 2011) UN Doc CCPR/C/GC/34 [hereafter: General Comment No 34], para 34.

This textually narrow scope has proven divisive. Some experts maintain that the list of identity factors is exhaustive,[81] while human rights mechanisms, including the ECtHR and UN human rights experts, advocate for a more inclusive interpretation.[82] The Human Rights Committee, though not formally adopting a more inclusive stance, has emphasised the connection between Article 20 and Articles 24 and 26 of the ICCPR,[83] which prohibit discrimination on more comprehensive grounds.[84] The Committee has also criticised States for a lack of hate speech prohibitions regarding gender, sexuality and disability in a number of country-specific concluding observations,[85] suggesting an openness to expand protected categories.[86] The UN Special Rapporteur on Freedom of Expression reinforced this position, noting that 'given the expansion of protection worldwide' over time, the 'United Nations human rights standards offer broader protection against discrimination than that afforded through the focus in article 20 (2) on national, racial or religious hatred.'[87]

81 William Schabas, *U.N. Covenant on Civil and Political Rights: Nowak's CCPR Commentary* (3rd rev edn, NP Engel 2019) 585; UNGA, 'Report of the United Nations High Commissioner for Human Rights on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred' (11 January 2013) UN Doc HRC/22/17.Add.4 [hereafter: Rabat Plan of Action] (these 'restrictions must be formulated in a way that makes clear that its sole purpose is to protect individuals and communities belonging to ethnic, national or religious groups, holding specific beliefs or opinions, whether of a religious or other nature, from hostility, discrimination or violence, rather than to protect belief systems, religions or institutions as such from criticism').

82 *Vejdeland and Others v. Sweden* App no 1813/07 (ECtHR, 9 February 2012) para 55; UNGA 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (n 3) para 32; UNGA, 'Report of the Special Rapporteur on Minority Issues, Fernand de Varennes' (3 March 2021) UN Doc A/HRC/46/57, para 88; UN Strategy and Plan (n 8) 2; UNGA, 'Report of the Special Rapporteur on Freedom of Religion or Belief, Ahmed Shaheed' (6 March 2019) UN Doc A/HRC/40/58, para 26; UNGA, 'Report of the Special Rapporteur on Freedom of Religion or Belief, Heiner Bielefeldt' (29 December 2014) UN Doc A/HRC/28/66, para 34; Committee of Ministers, 'Recommendation CM/Rec (2022)16 of the Committee of Ministers to member States on combatting hate speech' (20 May 2022) CM/Rec (2022)16, para 11 (the Committee includes 'racist, xenophobic, sexist and LGBTI-phobic' threats and insults within the scope of expressions of hate speech that are subject to criminal liability.)

83 General Comment No 34 (n 80) para 26.

84 ICCPR (n 87) Article 2(1).

85 Elbahtimy (n 77) 106 referencing UN Human Rights Committee, 'Concluding Observations on the Fifth Periodic Report of Mongolia' (2011) UN Doc CCPR/C/MNG/CO/5, para 10; 'Concluding Observations on the Fifth Periodic Report of Mongolia' (2011) UN Doc CCPR/C/MNG/CO/5, para 9; 'Concluding Observations on the Sixth Periodic Report of Poland' (2010) UN Doc CCPR/C/POL/CO/6, para 8; 'Concluding Observations on the Third Periodic Report of the United States of America' (2006) UN Doc CCPR/C/USA/CO/3, para 25.

86 *Ibid*.

87 UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye' (9 October 2019) UN Doc A/74/486, para 9.

In relation to Article 4 ICERD, an inter-conventional interpretation also broadens the textually limited scope. 'Racial discrimination' in Article 1(1) ICERD includes 'any distinction, exclusion, restriction or preference based on race, colour, descent, or national or ethnic origin.'[88] In its course-altering General Recommendation 35, the ICERD affirmed broadening the scope, stating that 'racist hate speech can take many forms and is not confined to explicitly racial remarks'.[89] The Committee has throughout the years 'included all the specific speech forms […] directed against groups, recognised in article 1 of the Convention' on grounds of 'race, colour, descent, or national or ethnic origin.'[90] This positions, for example, 'indigenous peoples, descent-based groups, and immigrant or non-citizens, including migrant domestic workers, refugees and asylum seeks, as well as speech directed against women members of these and other vulnerable groups,' within the protective reach of the Convention.[91] The Committee also indicated that islamophobia and antisemitism fall within the scope of Article 4, due to the overlap between religion and ethnicity,[92] reflecting an awareness of 'racialised communities' at the intersection between religion and race. However, it has thus far not applied this understanding in practice.[93]

Discriminatory disinformation predominantly targets vulnerable and marginalised communities. The referenced examples of discriminatory information (5.2.1) targeting racial, ethnic, or national identity – such as anti-black communities, anti-Asian stereotyping during the COVID-19 pandemic, or false claims about migrants – uncontroversially come within the ambit of both provisions. For other forms, including gendered or ageist disinformation, this depends on the acceptance of wider interpretation. The UN Special Rapporteur on Freedom of Expression's report on gendered disinformation did exemplify this application, concluding that Article 20(2) ICCPR, 'is considered to extend to sex and gender, on the basis of the principles of gender equality and non-discrimination enshrined in international law' and that '[t]hose forms of gender discrimination that meet the criteria set out in this provision are prohibited.'[94] This reasoning follows the harmonious interpretation connecting Article 20(2) with Articles 2(1) and 26, and parallels the ICERD approach of recognising 'racialised communities' at the intersection of protected characteristics. Thus,

88 IICERD (n 72) Article 1(1).

89 UN Committee on the Elimination of Racial Discrimination, 'General Recommendation No. 35: Combatting Racist Hate Speech' (26 September 2013) UN Doc ICERD/C/GC/35 (ICERD General Recommendation 35) para 7.

90 *Ibid.* para 6, 13(b).

91 *Ibid.* para 6.

92 *Ibid.*; Patrick Thornberry, *The International Convention on the Elimination of All Forms of Racial Discrimination* (Oxford University Press 2016) 283-285.

93 *Ibid.* 284.

94 UNGA 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Kahn' (n 3) para 32.

while traditional interpretations may have excluded certain forms of discriminatory disinformation, evolving interpretative approaches increasingly expand the protective scope of international law.

5.4.2 Advocacy and Intent

Advocacy for hatred constitutes the first substantive element of Article 20(2) ICCPR, though disagreement persists as to the scope of expressions qualifying as 'advocacy'. The *travaux préparatoires* of the ICCPR provides minimal guidance; during the drafting process of the provision, some delegates equated 'advocacy' with 'systematic and persistent propaganda', aligning it with the terminology in Article 20(1), while others characterised it as 'repeated and insistent expression', limiting 'advocacy' to hate speech campaigns of a certain scale.[95]

The Human Rights Committee has not provided the needed clarity. Its most substantive attempt was a proposed definition during the drafting process of the General Comment on the freedom of expression, initially defining 'advocacy' as 'public forms of expression that are intended to elicit action or response.'[96] However, neither this nor any other definition was retained in the final version of the General Comment or the 1983 General Comment on Article 20 ICCPR. Only the element that the expression should be public in nature consistently resurfaced in relevant case law by the Committee.[97]

Attempting to fill the gap, the UN Special Rapporteur on Freedom of Expression in 2012 characterised 'advocacy' as 'explicit, intentional, public and active support and promotion of hatred towards the target group' drawing upon the Camden Principles on Freedom of Expression and Equality.[98] Subsequent Rapporteurs endorsed this definition, while also introducing the 2019 Rabat Plan of Action' interpretation that 'advocacy is to be understood as requiring an intention to promote hatred towards the target group', which was taken *verbatim* from the Camden Principles.[99] These principles on 'Freedom of Expression and Equality', developed between 2008 and 2009 by

95 Temperman (n 60) 169.

96 UN Human Rights Committee, 'Draft General Comment 34: Article 19' (2nd rev draft, 28 June 2010) UN Doc CCPR/C/GC/34/CRP.3, para 53.

97 Temperman (n 60) 169 ft. 35; UN Human Rights Committee, *JRT and the WG Party v. Canada* Communication 104/1981, Decision of 6 April 1983, para 8(b).

98 UNGA 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (n 17) para 44(b); European Commission against Racism and Intolerance (ECRI), 'ECRI General Policy Recommendation on Combatting Hate Speech' (8 December 2015) Council of Europe, para 7(a).

99 UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (n 87) para 13; Rabat Plan of Action (n 81) para 21 ft. 5.

ARTICLE19, have subsequently taken a central position in academic and policy discussions on regulating harmful speech.[100]

Despite the scarcity of guidance, three features, or qualifiers, of 'advocacy' have crystallised over time. First, the speech must be public in nature; Article 20(2) ICCPR excludes statements made in an exclusively private setting, an interpretation that extends to 'incitement'.[101] The requirement of publicity refers to the context in which it was made rather than the actor – private actors can 'commit' public advocacy, through any medium. In *JRT and the WG Party v. Canada*, the Human Rights Committee found that the dissemination of antisemitic messaging by telephonic means 'clearly constitute the advocacy of racial or religious hatred' under Article 20(2).[102] In *Malcolm Ross v. Canada*, the Committee arrived at a similar conclusion regarding anti-Semitic pamphlets disseminated by a teacher at schools.[103] The critical question in relation to the emergence of new information and communication technologies is, however, not whether the expression must be public, but whether new platforms and communications can be considered 'public.'

Second, an intent requirement occurs in most authoritative interpretations. Nowak's ICCPR Commentary cautiously notes that the term 'advocacy' does seem 'to connote an element of motive', an arguably 'more suitable term than intent.[104] A 2001 Joint Statement by international experts indeed concluded that 'no one should be penalized for the dissemination of 'hate speech' unless it has been shown that they did so with the intention of inciting discrimination, hostility or violence.'[105] Almost 20 years later, thirty UN human rights experts expressed support for the Rabat Plan of Action standards, endorsing the definition of 'advocacy' as 'requiring an intention to promote hatred'[106] – a position supported by scholars and adopted in national legislation.[107]

100 ARTICLE19, 'The Camden Principles on Freedom of Expression and Equality' (April 2009).

101 Section 5.4.4 'Incitement'.

102 *JRT and the WG Party v Canada*, (n 97) para 8(b).

103 UN Human Rights Committee, *Malcolm Ross v Canada* Communication No 736/1997 (18 October 2000) UN Doc CCPR/C/70/D/736/1997, paras 6.2, 6.3, 7.2.

104 Wilson (n 54) 47; Schabas (n 81) 585; Oppositely, some courts and tribunal have argued that motive – as a distinct concept compared to intent – is not relevant, in Paul Behrens, 'Genocide and the Question of Motives' (2012) 10 Journal of International Criminal Justice 501, 508.

105 Organisation for Security and Co-operation in Europe, 'Joint Statement on Racism and the Media by the UN Special Rapporteur on Freedom of Opinion and Expression, the OSCE Representatives on Freedom of the Media and the OAS Special Rapporteur on Freedom of Expression' (27 February 2001) 2.

106 United Nations, 'Joint Open Letter on Concerns About the Global Increase In Hate Speech' (23 September 2019).

107 Temperman (n 60) 234; Jeroen Temperman and Evelyn Aswad, 'To Ban or Not to Ban Blasphemous Videos' (2013) 44 Georgetown Journal of International Law 4, 1313-1328.

Regarding the appropriate intent standard or threshold,[108] Temperman argues that 'anything less than a criminal intent requirement could jeopardise genuine research and journalism, and thus chilling free speech.'[109] The same applies to protecting forms of provocative art or satire, and unlike 'merely' offensive statements which may 'unconsciously' offend someone or to disseminate hateful messages', 'to *advocate hatred* is not something one does 'unconsciously and unwillingly.'[110] While hatred may be fuelled by expressions driven by naivety, negligence or recklessness on the side of the speaker, the juxtaposition of 'advocacy' and 'hatred' suggests a higher mental culpability threshold.[111] The actor engaging in the advocacy should thus have the intention to target an individual or group based on their protected identity characteristics, alongside the intent to advocate hatred against them.[112]

Third, advocacy arguably implies a certain scale of dissemination. Commentators have linked the notion of 'advocacy' to 'a minimal degree of dissemination,'[113] or 'a certain degree of intensity',[114] as otherwise it cannot rise to the level of incitement. This raises the question whether a one-off instance of discriminatory disinformation would be *a priori* excluded from the scope of Article 20 ICCPR. While advocacy's ordinary meaning does not suggest exclusion, it ultimately depends on whether 'advocacy', like incitement, is recognised as a continuing process in the legal context (5.4.4).

Publicity and the degree of dissemination or intensity do not pose any unique questions in the context of disinformation.[115] The intent component, however, excludes some forms of disinformation from the scope of unlawful hate speech. Although disinformation is intrinsically characterised by an intent to cause harm, discriminatory disinformation is not necessarily driven by the intent to advocate hatred, as established in section 5.2.1.2. Is hateful disinformation that is principally driven by monetary gain *a priori* excluded from the scope of Article 20(2)? Purposely exacerbating, *inter alia,* racialised and xenophobic disinformation for financial gain – to help sustain human traffick-

108 The *travaux préparatoires* of Article 20(2) leaves the matter of intent undecided; it was not fervently opposed, nor received substantial support, in Temperman (n 60) 208.

109 Temperman (n 60) 211.

110 *Ibid.* 210.

111 Ferreira Santos (n 25) 130; Temperman (n 60) 210; Rabat Plan of Action (n 81) 11 ('negligence and recklessness are not sufficient for an act to be an offence under article 20 of the Covenant, as this article provides for "advocacy" and "incitement" rather than the mere distribution or circulation of material').

112 Temperman (n 60) 211-212; ARTICLE19, 'Prohibiting Incitement to Discrimination, Hostility or Violence' (Policy Brief 2012) 22.

113 *Ibid.* 171.

114 McGonagle (n 7) 272.

115 Section 3.2.2 'Third Party Communication'; section 4.4.3 'Incitement to Terrorism'; section 5.5.2.1 'Public Nature'.

ing or actively work against abolitionist movements in systems of modern slavery –[116] likely constitutes advocacy for hatred.

While it may not be the principal intent of the actor behind it, such perpetrators are undoubtedly aware of the hateful nature of their conduct – beyond naivety, negligence or recklessness. Whether it, however, rises to the level of incitement is highly contextual. These extremities may suggest the need to lower the *mens rea* standard to knowledge or even negligence. However, as Temperman rightly highlights, '[t]he laws task should be limited to allocating criminal culpability to those who actually 'advocate' hatred that amounts to incitement.'[117] To protect legitimate scope, the application of 'advocacy' to discriminatory disinformation must be limited to expressions that are principally motived by the intent to advocate hatred.

5.4.3 Hatred

The element of 'hatred', alongside 'incitement', is the most in need of definitional clarity.[118] Featuring in both Article 20(2) and Article 4 ICERD – which obliges States to 'condemn all propaganda' that promotes 'racial hatred' and impose an obligation to 'declare as an offence' 'all dissemination of ideas based on […] hatred'[119] – neither framework has adopted a comprehensive definition. The Committee on the Elimination of Racial Discrimination (CERD) and the Human Right Committee have only occasionally accepted that statements amounted to or advocated 'hatred', yet without providing instructive clarification on these determinations.[120]

More than with defining 'advocacy', capturing 'hatred' as an affective state in legal terms is challenging.[121] On the spectrum of negative expressions, the threshold from expressing emotion to becoming hatred under international law is ambiguous.[122] An early draft of the General Comment No 34 on Freedom of Expression attempted clarity, defining 'hatred' in line with its ordinary meaning as 'intense emotions of opprobrium, enmity and detestation towards

116 Amazeen *et al.* (n 30) 24; Mitchell (n 30); Patrick Wintour, "Fake news': Libya Seizes on Trump Tweet to Discredit CNN Slavery Report' The Guardian (38 November 2017); Jaya Prakash *et al.*, 'Human Trafficking and the Growing Malady of Disinformation' (2022) 10 Public Health 987159.

117 Temperman (n 60) 211.

118 Toby Mendel, 'Does International Law Provide for Consistent Rules on Hate Speech?' in Michael Herz and Peter Molnar, *The Content and Context of Hate Speech* (Cambridge University Press 2012) 423.

119 IICERD (n 72) Article 4.

120 *JRT and the WG Party v Canada* (n 97) para 8(b); *Malcolm Ross v. Canada* (n 103) para 11.5.

121 McGonagle (n 7) 273; Temperman (n 60) 174.

122 Mendel (n 119) 427.

a target individual or group.'[123] Though ultimately deleted from the final version, the wording re-occurred in the influential Camden Principles and the Rabat Plan on Hate Speech.[124] The UN Special Rapporteur on Freedom of Expression similarly characterised 'hatred' as 'a state of mind characterized as intense and irrational emotions of opprobrium, enmity and detestation towards the target group' emphasising the 'severity of hatred' to reiterate that advocacy should amount to 'the most severe and deeply felt form of opprobrium' to qualify as prohibited forms of expression.[125] Whether this threshold is reached depends on 'the severity of what is said, the harm advocated, magnitude and intensity in terms of frequency, choice of media, reach and extent.'[126]

Scholarly interpretations have developed along similar lines. McGonagle, for example, cites Partsch's 'commendable attempt' in describing hatred as 'an active dislike, a feeling of antipathy or enmity connected with a disposition to injure',[127] while Elbathimy, in her work on the right to be protected from incitement, links 'hatred' to the content of expressions covered by the provision.[128] Though States generally fail to provide the necessary clarity, the Canadian Supreme provided a comprehensive account in *R v. Keegstra*, stating that:

> '[h]atred is predicated on destruction, and hatred against identifiable groups and therefore thrives on insensitivity, bigotry and destruction of both the target group and of the values of our society. Hatred in this sense is a most extreme emotion that belies reason; an emotion that, if exercised against members of an identifiable group, implies that those individuals are to be despised, scorned, denied respect and made subject to ill-treatment on the basis of group affiliation.'[129]

Either way, 'hatred' restricts prohibited 'advocacy' to expressions characterised by extreme emotions or feelings, excluding weaker, less inimical forms of hate speech. Yet, as Temperman observes in this reconstruction of the *actus reus* of Article 20(2) ICCPR, the emphasis on emotional states complicates address-

123 Schabas (n 81) 586; Draft General Comment No 34 (n 96) para 53; Oxford English Dictionary defines 'hatred' as '[a] feeling of intense dislike or aversion towards a person or thing; an emotion in which such a feeling is experienced; loathing; hostility; malevolence', Accessed 11 January 2025.

124 Although the Rabat Plan equated hatred and hostility: 'the term "hatred" and "hostility" refer to intense and irrational emotions of opprobrium, enmity and detestation towards the target group', in Rabat Plan of Action (n 81) ft. 5

125 UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (n 17) para 44.

126 *Ibid.* para 45.

127 McGonagle (n 7) 273 citing Josef Partsch, 'Freedom of Conscience and Expression, and Political Freedoms' in Louis Henkin (ed), *The International Bill of Rights: The Covenant on Civil and Political Rights* (Colombia University Press 1981) 228.

128 Elbahtimy (n 77) 107.

129 *R v Keegstra* [1990] 3 SCR 697, Part VII(D)(i) (Supreme Court of Canada).

ing extreme speech that is not explicitly accompanied by such emotions. It is, he argues, 'not unimaginable that a person would incite the most odious acts – including crimes against humanity – ostensibly without much emotions at all.'[130] While 'hatred' as an affective state complicates the legal delineation of both provisions, it illustrates significant convergence between discriminatory disinformation and prohibited hate speech.

Discriminatory disinformation may not always *advocate* hatred, but it does contribute to an atmosphere of hatred and normalise its spread.[131] Such an environment, some delegates argued during the drafting of the ICERD, 'inevitably lead[s]' to discrimination.[132] Discriminatory disinformation's persistence stems precisely from its exploitation and amplification of hatred or hatred-adjacent emotions: fear, anger and resentment.[133] While contributing to hatred does not necessarily equal advocacy, it does consistently contribute to an atmosphere of hatred in which expressions are more likely to cross the threshold into advocacy and subsequent incitement.[134]

5.4.4 Incitement

'Incitement' represents the pivotal concept in both Article 20(2) ICCPR and Article 4 ICERD. Under Article 20(2), 'advocacy for hatred' is prohibited only when it 'constitutes incitement', while Article 4 ICERD restricts expressions only if they amount to 'incitement' in the form of 'advocacy or threats'. Despite its centrality, both Conventions provide limited definitional clarity on the meaning of incitement, leading to varied interpretations across international bodies.

The Committee on the Elimination of Racial Discrimination notes that incitement 'characteristically seeks to influence others to engage in certain forms of conduct, including the commission of crime, though advocacy or

130 Temperman (n 60) 173.

131 Section 5.2.1 'Conceptualising Discriminatory Disinformation; Bilewicz and Soral (n 11) 3-33.

132 Natan Lerner, *The U.N. Convention on the Elimination of All Forms of Racial Discrimination* (Sijthoff & Noordhoff 1980) 52 citing E Ketrynzki, Statement to the UN Sub-Commission on Prevention of Discrimination and Protection of Minorities, UN Doc E/CN.4.Sub.2/SR.418.

133 Grambo (n 20) 1317-1318; UNGA, 'Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance' (n 47) para 25; Kelly Garett, Eric C Nisbet and Emily Lynch, 'Undermining the Corrective Effects of Media-Based Political Fact-Checking? The Role of Textual Cues and Naïve Theory' (2013) 63 Journal of Communication 617, 610; DJ Flynn, Brendan Nyhan and Jason Reiflet, 'The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs about Politics' (2017) 38 Advances in Political Psychology 1, 133; UNGA, 'Report of the Special Rapporteur on Contemporary Forms of Racism, Racial Discrimination, Xenophobia and Related Intolerance' (n 8) para 20.

134 *Ibid.*

threats,'[135] emphasising the intent and the ability to mobilise the audience.[136] The earlier mentioned draft of General Comment 34 defined incitement as 'the need for the advocacy to be likely to trigger imminent acts of discrimination, hostility or violence against a specific individual group.'[137] Though ultimately rejected, subsequent definitions proposed by, *inter alia,* the UN Special Rapporteur on Freedom of Expression, echo this understanding, opining in 2012 that 'incitement refers to statements about national, racial or religious groups which create an imminent risk of discrimination, hostility or violence against persons belonging to those groups.'[138] Taken from the Camden Principles, this definition was later adopted in the Rabat Action Plan and reiterated in subsequent reports by UN Special Rapporteurs. Applied to disinformation, the UNSP's unfortunately did not further clarify the term, qualifying 'incitement' as 'advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility, violence (commonly referred to as "hate speech"), war crimes, crimes against humanity and genocide.'[139]

The ICERD's General Recommendation No 35 provides a more comprehensive framework for evaluating whether 'the dissemination of ideas' under Article 4 ICERD qualifies as incitement. It outlines five evaluative factors adapted from the Rabat Plan: the content and form of speech; the economic, social, and political climate; the position or status of the speaker; the reach of speech; and its objective.[140] Additionally, the Recommendation emphasises the need to consider 'the intention of the speaker, and the imminent risk or likelihood that the conduct desired or intended by the speaker will result from the speech in question […].'[141]

Synthesising these interpretations, incitement under both Article 20(2) ICCPR and Article 4 ICERD is characterised by 1) an intent to mobilise the audience to action; and 2) a likelihood or imminent risk that the conduct the

135 ICERD General Recommendation 35 (n 105) para 16.

136 Temperman (n 60) 180-181, UN Human Rights Committee, *Faurisson v. France*, Communication No 550/1993, Concurring Opinion Rajsoomer Lalah (8 November 1996) UN Doc CCPR/c/58/1993, para 9 ('propagated ideas tending to revive Nazi doctrine and the policy of racial discrimination' constituted 'incitement, at the very least, hostility and discrimination', because the statements 'were […] found to have been of such a nature as to raise or strengthen antisemitic feelings' in others).

137 Draft General Comment No 34 (n 96) para 53.

138 UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (n 17) para 44; UNGA, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (n 87) para 16; Camden Principles (n 100) 10.

139 UNGA, 'Disinformation and Freedom of Opinion and Expression During Armed Conflicts: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (n 3) para 16.

140 ICERD General Recommendation 35 (n 89) para 15.

141 *Ibid.* para 16; Lerner (n 132) 52 (aligning with the discussed meaning of 'incitement' during the drafting of the Convention as a 'conscious and motivated act').

speaker aims to trigger will occur.[142] A third requirement is that the incitement must be public,[143] which refers to the location and context of the expression, not the nature of the actor involved.[144] The same points of contention on the application of traditional notion of 'public' to the online realm exist as in relation to defamatory speech (3.2.2), incitement to terrorism (4.4.3) and incitement to commit genocide (5.5.2.1).

The distinctive features of discriminatory disinformation – its deceptive nature, technological amplification, and cumulative cognitive exploitation – require a recalibration of traditional incitement assessments. The core requirements of incitement (intent to mobilize, likelihood of resulting harm, and public expression) must be re-interpreted to comprehend these operational realities, enabling meaningful analysis of whether and when disinformation amounts to incitement. The following assessment demonstrates how discriminatory disinformation systematically influences each contextual factor used to evaluate incitement. Whether discriminatory disinformation itself qualifies as incitement is contingent upon a contextual analysis of:

1. The likelihood or imminent risk that discriminatory disinformation results in the proscribed harm (5.4.4.1)
2. The content, form, and presentation: the specific nature of the information, including its linguistic features, visual elements, and mode of delivery (5.4.4.2)
3. The economic, social and political climate: the prevailing conditions at the time of the audience's exposure to the messages may influence their receptivity and potential reactions (5.4.4.3)
4. The speaker, audience, and technological factors: the identity and status of the speaker, the characteristics of the target audience, and the technological means employed in the production and dissemination of the information (5.4.4.4).

142 Temperman (n 60) 182-183; Thornberry (n 92) 293; Evelyn Aswad and David Kaye, 'Convergence and Conflict: Reflections on Global and Regional Human Right Standards on Hate Speech' (2022) 20 Northwestern Journal of Human Rights 3,177-179; Devin Carpenter, 'So Made That I Cannot Believe: The ICCPR and the Protection of Non-Religious Expression in Predominantly Religious Countries' (2017) 18 Chicago Journal of International Law 1, 240-241.

143 Schabas (n 81) 584.

144 Thornberry (n 92) 291 (the ICERD Committee 'has been critical of criminal provisions on 'dissemination' that define the prohibition to dissemination among the public'); UN Human Rights Committee, *Gelle v Denmark*, Communication No. 34/2004 (6 March 2006) UN Doc ICERD/C/68/D/34/2004, para 6.5; *Jama v Denmark*, Communication No. 41/2008 (21 August 2009) UN Doc ICERD/C/75/D/41/2008, para 6.5.

5.4.4.1 Likelihood or Imminent Risk

Hateful incitement functions as an inchoate offense.[145] The preventive purpose of both Article 20 ICCPR and Article 4 ICERD necessitates intervention before the proscribed harms materialize, while simultaneously requiring a plausible expectation that harm could occur. The qualification as incitement hinges on the risk and likelihood of harm resulting from expressions.[146]

ICERD General Recommendation 35, however, introduced a noteworthy distinction between 'an imminent risk *or* likelihood (emphasis added)' of resulting harm.[147] This distinction likely signals a more nuanced approach to evaluating the potential consequences of inciting speech in terms of geographical and temporal proximity between expression and subsequent harm. An *imminent* risk implies a close temporal and geographical proximity between the speech and harm to the target group or individual(s), exemplified by the United States's First Amendment standard requiring speech to incite 'imminent lawless action'.[148] In contrast, a *likelihood* of harm, without the qualifier of imminence, allows for a broader interpretation encompassing expressions likely to lead to the proscribed harms over extended periods of time or through cumulative effects of multiple speech acts. *Faurisson v. France* illustrates this approach, where the Human Rights Committee determined that statements 'of a nature as to raise or strengthen antisemitic feelings' could constitute incitement to discrimination without an immediate threat of harm.[149]

This distinction is relevant to modern information manipulation threats, especially discriminatory disinformation, which causes harm gradually rather than inciting immediate violence. While real and potentially foreseeable, the risks are often temporally distant from the initial dissemination of false or misleading information. Prior discussed research demonstrates that cognitive and behavioural processes exploited by discriminatory disinformation leading

145 ICERD General Recommendation 35 (n 89) para 16; Temperman (n 60) 183; UN Human Rights Council, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue: Addendum' (29 May 2012) A/HRC/22/17/Add.4, para 29(f); UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue' (n 17) para 45(e).

146 The standards are mostly used interchangeably or are merged; Temperman, *inter alia*, refers to 'a likelihood of (imminent) harm' in Temperman (n 60) 183, and Parmar speaks of 'likelihood, including imminence' in Sejal Parmar, 'The Rabat Plan of Action: A Critical Turning Point in International Law of "Hate Speech"' in Peter Molnar (ed), *Free Speech and Censorship Around the Globe* (Central European University Press 2015) 230.

147 ICERD General Recommendation 35 (n 89) para 16; ECRI (n 98) para 14 ('the element of incitement entails there being either a clear intention to bring about the commission of acts of violence, intimidation, hostility or discrimination or an imminent risk of such acts occurring as a consequence of the particular hate speech used').

148 Brandenburg v Ohio 395 US 444 (1969) (United States Supreme Court).

149 *Faurisson v. France* (n 136) para 9.6.

to changes in audience behaviour – the ultimate measure of incitement's impact – typically unfold over prolonged periods rather than instantly.[150] Systematic discrimination, racist attacks, and hate crimes rarely result from single, isolated, utterances, but emerge from sustained exposure to harmful rhetoric. This cumulative effect aligns more closely with a likelihood standard than an imminent risk standard.

Consequently, different proscribed harms thus demand different thresholds of risk and proximity. While incitement to physical violence might link more directly to specific triggering expressions, the preceding discriminatory beliefs and practices often result from cumulative exposure to various expressions over time. In other words, incitement to discrimination needs to be evaluated differently than incitement to violence. Domestic legal approaches support such a differentiated approach. Criminal sanctions for incitement to violence are generally higher than those for incitement to discrimination,[151] which could justify a stringent threshold (imminent risk) for violence and more flexible threshold (likelihood) for discrimination or hostility. Adopting this differentiation in international standard setting could resolve key definitional complexities surrounding both provisions, enabling a more nuanced recognition of the diverse types of speech that constitute incitement.

This nuanced standard must not be mistaken for lowering the threshold of incitement. The fundamental assessment remains the same, centralising the question whether at the time the expressions were made the risk of harmful acts to be perpetrated by the speech act's target group existed.[152] Moving beyond the stringent imminence standard acknowledges the reality of the gradual, cumulative nature of harm resulting from sustained exposure to discriminatory content, disinformation or otherwise. It also enables more meaningful consideration of the contextual factors that, *inter alia,* the ICERD deems significant.[153] By adopting a more realistic standard, legal systems can better anticipate and respond to the evolving landscape of information manipulation, while maintaining necessary freedom of expression protections. This differentiated approach to risk assessment also guides the evaluation of the content, form, and presentation of discriminatory disinformation as potentially inciting.

5.4.4.2 Content, Form and Presentation

The substantive overlap between hate speech and disinformation – identified in section 5.2.1.1 – demonstrates that discriminatory disinformation employs dehumanising, stigmatising, stereotyping and/or demonising language –

150 Section 5.2.2 'Cognitive and Behavioural Mechanisms of Discriminatory Disinformation'.
151 Clooney and Gardoll (n 75) 156-171.
152 Temperman (n 60) 182.
153 ICERD General Recommendation 35 (n 89) para 14.

mirroring recognised patterns in unlawful hate speech.[154] A textual analysis of such content may uncover direct calls for violence, identify targeted groups or individuals and the grounds for such targeting. Crucially, content combined with tone and presentation can indicate the speaker's intent. These three elements – content, tone and style/presentation – serve as contextual factors recognised by international human rights mechanisms and authoritative policy strategies.[155]

The emphasis on tone and form, alongside the content,[156] addresses incitement's relational nature. Ultimately, how the audience interprets, internalises and reacts to a message determines its (potential) impact. Linguistically provocative and direct speech more likely triggers an emotional response and imminent action.[157] Different types of content and form may therefore reveal the motive or intent of speakers, indicating their 'underlying or motivating feelings of opprobrium, enmity and detestation.'[158] For example, the spread of falsities is considered an indicator of malevolent intent, while artistic expression or satire suggest benevolent intent.[159]

Several format and language characteristics of discriminatory disinformation signal prohibited incitement. Content, *inter alia*, mimicking legitimate news articles, reports, or scientific publications exploits established credibility norms to amplify discriminatory messages while creating resistance to corrections from imitated authorities and institutes. This "fake news" approach and instrumentalisation of scientific disinformation weaponises perceived legitimacy to enhance persuasiveness – directly implicating the 'likelihood' standard. In other words, when false information appears credible through its presentation, it more readily satisfies legal standards requiring foreseeable harm, even without explicitly inciting language.

The medium and form of presentation additionally influence this assessment. The potential of visual discriminatory disinformation to incite warrants heightened scrutiny; visual hate speech exerts substantially greater influence

154 Rabat Plan of Action (n 81) para 22; ICERD General Recommendation 35 (n 89) para 15; UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (n 17) para 45(c) (an assessment of the content includes 'form, style, nature of the arguments deployed in the speech, magnitude or intensity of the speech, […] and the degree to which the speech is provocative or direct').

155 *Ibid*.

156 The Rabat Plan of Action positions it as 'one of the key foci of the court's deliberations and is a critical element of incitement', in Rabat Plan of Action (n 81) 11.

157 Section 5.2.2 'Cognitive and Behavioural Mechanisms of Discriminatory Disinformation'; Rabat Plan of Action (n 81) para 22; ICERD General Recommendation 35 (n 89) para 15.

158 Temperman (n 60) 185; section 5.5.2.2 'Direct Incitement'.

159 Robert Post, 'Hate Speech' in Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (Oxford University Press 2009) 127, 133-135.

than textual forms,[160] creating more immediate emotional responses and stronger memory imprints. This impact suggests that visual discriminatory disinformation – especially synthetically generated content utilising deepfake audio-visual technology – more readily meets incitement thresholds under both Article 20(2) ICCPR and Article 4 ICERD. Discriminatory disinformation typically employs linguistically subtle, more manipulative language gradually 'poisoning the mind', although this indirectness does not necessarily diminish its effectiveness in inciting violence or other proscribed harms.[161] Equally – or more – dangerous is overtly provocative and inflammatory discriminatory disinformation, as exemplified by the extensively documented anti-Asian discriminatory disinformation during the COVID-19 pandemic.[162] Consequently, the spectrum from subtle to overt forms necessitates a nuanced consideration of both the immediate impact of discriminatory disinformation's content, form and presentation as well its cumulative effects to determine whether it amounts to prohibited incitement. In this evaluation, the impact depends on the broader economic, social, and political climate in which speech circulates – a context that discriminatory disinformation itself helps construct and distort over time.

5.4.4.3 Economic, Social and Political Climate

The economic, social, and political climate influences how content, form and presentation are perceived and thus whether discriminatory disinformation is likely to incite discrimination, hostility or violence. The ICERD has formalised and emphasised this contextual factor, recognising that statements that appear 'innocuous or neutral' in one context, may become inflammatory and dangerously significant in another.[163] The Committee has positioned the economic, social, and political climate 'prevalent at the time the speech was made and disseminated, including the existence of patterns of discrimination' as an influential elements of potential incitement.[164] Applied to disinformation, in its 2005 report 'indicators of systematic and massive racial discrimination', the ICERD already characterised propaganda as an 'important component[s] of situations leading to conflict and genocide,' and encouraged

160 Ursula Kristin Schmid *et al.*, 'How Social Media Users Perceive Different Forms of Online Hate Speech: A Qualitative Multi-Method Study' (2024) 26 New Media & Society 5, 2625; Anna Stefaniak and Mikolaj Winiewski, 'Differentiating Between Direct and Indirect Hate Crime: Results From Poland' (2022) 10 Journal of Social and Political Psychology 1, 86-105; Jessica Lin, 'Levering World Knowledge in Implicit Hate Speech Detection' (Georgetown University, Department of Linguistic, Working paper December 2022).

161 Bianca Cepollaro *et al.*, 'Slurs in Quarantine' (2023) 39 Mind & Language 381-396.

162 Jae Yeon Kim, 'Misinformation and Hate Speech: The Case of Anti-Asian Hate Speech During the COVID-19 Pandemic' (2021) Journal of Online Trust and Safety, 1-14.

163 ICERD General Recommendation 35 (n 89) para 15.

164 *Ibid*.

States to combat racist and xenophobic propaganda as a 'contemporary form of racial discrimination.'[165] Here, the Committee identified two contextual risk indicators in particular:

1. The 'systematic and widespread use and acceptance of speech or propaganda promoting hatred and/or inciting violence against minority groups, particularly in the media.'
2. 'Grave statements by political leaders/prominent people that express support for affirmation of superiority of a race or an ethnic group, dehumanize and demonize minorities, or condone or justify violence against a minority.'[166]

Both patterns are well-recognised mechanism and strategies used in contemporary discriminatory disinformation – as substantiated in chapter one and section 5.2.1. The deliberate and systematic dissemination characterising disinformation intensifies the propaganda concerns raised, presenting several implications for legal analysis. First, if the widespread use of propaganda promoting hatred is an indicator for systematic discrimination and violence, discriminatory disinformation constitutes an elevated threat. Second, if systematic and widespread use of traditional media is 'particularly' worrying,[167] disinformation's penetration of online and social media, and algorithmic amplification represents an aggravating factor under ICERD's framework. And third, the increasing deployment of false narratives by 'political leaders/prominent people' further exemplifies how discriminatory disinformation intensifies patterns ICERD recognises as precursors to incitement and violence.[168]

Inclusion of the economic, social, and political climate in incitement assessment positions the findings from section 5.2.1.4 and 5.2.2.2 as an integral part of legal analysis. Even discriminatory disinformation that in itself does not amount to unlawful hate speech contributes to a social and political climate where explicit incitement finds receptive audiences – a mechanism consistent with the cumulative effect identified in Section 5.2.2. The discriminatory 'information climate' exacerbates 'existing patterns of discrimination' – a factor centralised by the ICERD.[169] This implies that discriminatory disinformation constitutes a pervasive indicator of a climate that is conducive to specific incitement cases and overall heightened risks of systematic and racial discrimination. Finally, as argued in section 5.2.1.4, the relationship between disin-

165 UNGA 'Official Records of the General Assembly, Sixtieth Session, Supplement No. 18' (2005) UN Doc A/60/18, chpter II, para 19-20.

166 *Ibid.*

167 *Ibid.*

168 Jace Valcore *et al.*, '"We're Led by Stupid People": Exploring Trump's Use of Denigrating and Deprecating Speech to Promote Hatred and Violence' (2023) 80 Crime, Law and Social Change 237-256; Tom Phillips, 'Jair Bolsonaro's Racist Comments Sparks Outrage From Indigenous Groups' The Guardian (24 January 2020).

169 ICERD General Recommendation 35 (n 89) para 15.

formation and discrimination is two directional: discriminatory disinformation both shapes and is shaped by existing patterns of discrimination. This dynamic becomes particularly significant when considering the third contextual factor: the relationship between speakers, audiences, and the technological infrastructure that mediates their interactions.

5.4.4.4 Speaker, Audience and Technology

The final contextual factor in incitement analysis – the relationship between speakers and audiences – is transformed by discriminatory disinformation. Traditional incitement assessments focus on a speaker's ability to exert influence over an audience, typically derived from formal position or status, and geographical proximity. Influence implies a capacity on behalf of the speaker vis-à-vis the audience.[170] On the receiving end, the audience's susceptibility to incitement is determined by an interplay of personal, inter-group, and societal circumstances, which can either increase vulnerability or contribute to resilience against inciting messages.[171] Discriminatory disinformation manipulates this susceptibility through falsely manufacturing a perception of expertise or authority; exploiting audience cognitive vulnerabilities through targeted deception; and leveraging technological infrastructure to create unprecedented reach and impact.

Human rights mechanisms consistently identify political figures and other public opinion-formers as particularly influential speakers.[172] Contemporary discriminatory disinformation discourse subscribes to this observation. Donald Trump's reiteration of conspiracy theories about COVID-19's origin and his insistence on the use of the term 'Chinese virus' capitalised on existing animosity towards immigrants of Asian descent, correlating with a spur in hate crimes.[173] As head of State and head of government, his extensive endorsement of anti-Muslim rhetoric similarly produced demonstrable stigmatisation and incited societal exclusion.[174] In Tunesia, Kais Saied promoted racist conspiracy theories about Sub-Saharan Africans,[175] and in Poland, 2019 marked the year when the incumbent president Andrzej Duda instrumentalised false

170 *Ibid.* para 15; Rabat Plan of Action (n 81) para 29.
171 Gelber (n 6) 393-414.
172 ICERD General Recommendation 35 (n 89) para 15.
173 Brendan Lantz *et al.*, 'Fear, Political Legitimization, and Racism: Examining Anti-Asian Xenophobia during the COVID-19 Pandemic' (2022) 13 Race and Justice 1; Mervat Ahmed, 'Polarization and Negative-Other 'China' Presentation in US President Trump's COVID-19 Tweets: A Critical Discourse Analysis' (2021) Cairo Studies in English: Journal of Research in Literature, Linguistics and Translation Studies 2, 150-151.
174 Angela Hefti and Laura Ausserlandscheider, 'From Hate Speech to Incitement to Genocide: The Role of the Media in the Rwandan Genocide' (2020) 38 Boston University International Law Journal 1, 28.
175 Africa Centre for Strategic Studies (n 30) Infographic.

narratives to demonise the LGBTQ community.[176] In the Netherlands, politician Geert Wilders was found guilty of incitement to discrimination for systematically disseminating incorrect statistics and making demeaning statements about migrant communities, particularly those of Moroccan descent.

The landscape of influential speakers has, however, expanded beyond traditional political figures. A growing number of influential individuals online, ranging from former comedians to wellness gurus and failed documentary makers,[177] contributes to the pervasiveness of inciting discriminatory disinformation. This development illustrates that the position or status of speakers is always relative to their audience; these figures attain their influential status online beyond the traditional stratagems and formal mechanisms of authority, power and influence.[178] They create online spheres of influence that operate outside conventional public scrutiny and democratic corrective mechanisms. These spheres are known for normalising intolerant and racist behaviour online and lowering the mental threshold for others to openly engage in hate speech on- and offline.[179]

Regarding the audience, factors of influence include the type of relationship between the audience and the speaker(s), the size of the audience,[180] as well as their susceptibility to the hateful advocacy. The manipulative, occasionally labelled 'coercive', nature of disinformation is designed to capitalise on cognitive vulnerabilities, as constructed in section 5.2.2. Exploitation extends to the perceived authority and legitimacy of the speaker – disinformation is often instrumentalised to attribute credibility, legitimacy and knowledge to a speaker without any grounding – as well as technological means of transmission that characterise the modern information dissemination broadly. Consequently, assessing the likelihood of incitement must prioritise the type of media involved, the reach of the speech, and the frequency of exposure to the message.[181] Acknowledging the changing communication dynamics, the ICERD emphasises repetition as an indicator of 'a deliberate strategy to engender hostility towards ethnic and racial groups' Beyond indirectly addressing

176 Anne Applebaum, 'Poland's Rulers Made Up a 'Rainbow Plague' The Atlantic (14 July 2020).

177 Including former comedian turned conspiracy thinker Joe Rogan (e.g. Joshua Cohen, 'Joe Rogan Provides A Platform To HIV/AIDS Denialists' Forbes (15 February 2024); documentary maker and conspiracy theorist Alex Jones (e.g. Juan A Lozano, 'Who is Alex Jones? The Conspiracist and Dietary Supplement Salesman Built an Empire Over Decades' AP News (15 June 2024)) and former comedian Russell Brand (e.g. Rachel Schraer, 'Russell Brand: How the comedian built his YouTube audience on half-truths' BBC News (21 September 2023).

178 Ishani Maitra, 'Subordinating Speech' in Ishani Maitra and Mary Kate McGowan (eds), *Speech & Harm: Controversies over Freech* (Oxford University Press 2012) 41-68.

179 Wilson (n 54) 231.

180 UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (n 17) para 45(d).

181 Wilson (n 54) 230.

disinformation's algorithmic dimension which facilitates repeated exposure, the Committee's reference to a 'deliberate strategy' highlights the convergence between disinformation and incitement. Not only does this increase the likelihood of escalation, the presence of a '*systematic campaign* of verbally attacking a target group' may, as Temperman notes, 'be circumstantial evidence of someone's intention to incite.'[182]

In sum, the evolving nature of online influence, combined with digital platforms' virtually unlimited reach indicates that discriminatory disinformation constitutes an aggravating or risk factor for all contextual factors determining incitement under international law. It exploits new spheres of influence between speakers and audiences, leverages technological amplification to maximise harm, and creates cumulative effects that traditional legal frameworks struggle to capture. While the existing indicators provide a foundation for addressing this threat as a form of prohibited incitement, there are several systemic misalignments. To mitigate these, the doctrine needs to modernise it understanding of authority and integrally consider "new" audience-speaker spheres of influence. The assessment must also more clearly account for the unique characteristics of digital communication environments that amplify discriminatory content.

5.4.5 Non-Physical Harm

Both Article 20(2) ICCPR and Article 4 ICERD prohibit advocacy that incites proscribed harms, including non-violent outcomes.[183] This recognition that hate speech can inflict significant damage without explicitly calling for violence broadens the applicability of this prohibition to discriminatory disinformation. Disinformation operates in the realm of intangible and indirect harms, reinforcing hostile attitudes and discriminatory practises, mostly without directly inciting violence. Scholars debate whether these provisions nevertheless establish a hierarchy among different forms of harm. Fino, *inter alia*, opines that 'there is an argument to be made that incitement to violence is graver than incitement to discrimination or hostility' since 'incitement to violence leading to physical or psychological harm to the person is more grievous than that leading to their discrimination or to hostility against them,' unless the latter amounts to 'psychological harm which is tantamount to violence.'[184]

182 Temperman (n 60) 216; *Mohamed Rabbae, ABS and NA v The Netherlands* (n 74) paras 10-11.

183 John T Bennett, 'The Harm in Hate Speech: A Critique of the Empirical and Legal bases of Hate Speech Regulation' (2016) 43 Hastings Constitutional Law Quarterly 445.

184 Audrey Fino, 'A Critique of the UN Strategy and Guidance on 'Hate Speech': Some Legal Considerations' (2023) 41 Netherlands Quarterly for Human Rights 211, 202; Dias (n 66) 402-404.

The inclusion of such hierarchy, however, finds little support in either the provision's text or drafting history. As Temperman observes, during the ICCPR drafting, delegates considered that 'despite the scale difference, […] 'incitement to hostility' and 'incitement to discrimination' deserved combatting as much as 'incitement to violence' does'.[185] The ICCPR's authoritative commentary affirms that 'the prohibition in Article 20 related to "incitement to discrimination, hostility *or* violence" […] literally means that incitement to discrimination without violence must also be prohibited' when constituting advocacy of hatred.[186] The relationship between these forms of harm also complicates any attempt at hierarchical categorisation. Section 5.2.2. illustrated how discrimination and hostility function as precursors to violence, suggesting they may ultimately cause equally serious consequences through their cumulative effects over time – a pattern particularly relevant to discriminatory disinformation's gradual impact.

While legitimate concerns exist that prohibiting incitement to non-violent harms could establish an unduly low threshold for speech restriction, the solution lies not in narrower the scope of the provisions or establishing a hierarchy. Rather, it requires developing clearer delineations and appropriate thresholds for unlawful incitement to discrimination and hostility. This reflects the Convention's intentions that creating an atmosphere of hatred constitutes legitimate grounds for restricting expression.[187] This approach simultaneously addresses the outstanding question whether expressions must expressly cause harm under one specific category or may more broadly qualify as 'discrimination, hostility or violence.'[188] Clearly delineating these harms encounters the earlier addressed difficulty of measuring intangible or emotional harm, both *a priori* and *ex post*, which unlike violence, 'cannot be detected or defined with medical objectivity'. Hence, 'the determination is therefore necessarily subjective'[189] as McGonagle rightly observes. The inclusion of the cognitive and behavioural mechanism into the concept of 'discriminatory disinformation', however, provides a promising avenue for mitigating this difficulty. Drawing from these findings, the analysis examines the contours of 'discrimination' (5.4.5.1) and 'hostility' (5.4.5.2) under Article 20(2) ICCPR, while also addressing Article 4 ICERD's prohibition on 'all dissemination of ideas based on racial superiority or hatred' which frequently encompasses discriminatory disinformation (5.4.5.3).

185 Temperman (n 60) 189.

186 Many States have indeed criminalized incitement to hatred accordingly, in Schabas (n 81) 576, 583-584.

187 Timmerman (n 50) 48 citing UN Doc A/C.3/SR.1079, para 9; Elbahtimy (n 77) 101.

188 Temperman (n 60) 187; Agnas Callamard, 'Combatting Discrimination and Intolerance with a Free Speech Framework' (2010) 5 Religion and Human Rights 153-169, 161.

189 McGonagle (n 7) 273.

5.4.5.1 Incitement to Discrimination

'Discrimination' is the first proscribed harm under Article 20 ICCPR, reflecting what Nowak describes 'as a special State obligation to take preventive measures at the horizontal level to enforce the right to […] equality'.[190] Among the three contingent harms in Article 20 ICCPR, 'discrimination' is the only one with a clear counterpart in the Convention. Article 4 ICERD retained a similar structure, requiring that States '[s]hall declare an offence punishable by law [….] incitement to discrimination' and 'prohibit organizations […] and all other propaganda activities, which promote and incite racial discrimination.' Similar to the ICCPR, throughout the drafting of the Convention, this prohibition's scope was considered among the most controversial points.[191]

The ordinary meaning of discrimination refers to behaviour, to an act that creates, sustains or reinforces advantages of one group and their members over another,[192] making a distinction, leading to exclusion, restriction or preference based on certain aspects of an individual's identity. The Human Rights Committee aligned itself with this approach, defining discrimination as 'any distinction, exclusion, restriction or preference which is based on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, and which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise by all persons, on an equal footing, of all rights and freedoms.'[193] Unless grounds for differentiating are 'reasonable and objective' and are in furtherance of 'a purpose which is legitimate under the [ICCPR] Covenant',[194] this behaviour is prohibited by international law.

This broad definition, however, creates challenges when applied to Article 20(2) as Nowak observes; '[i]t is most difficult to conceive of an advocacy for national, racial or religious hatred that does not simultaneously incite discrimination.'[195] Without an additional threshold – beyond the vague contours of incitement also constituting 'advocacy of hatred' – this interpretation undermines the overall credibility on the provision.[196] In response, scholars have

190 Schabas (n 81) 576; *Mohamed Rabbae, ABS and NA v The Netherlands* (n 74) para 9.7; General Comment No 34 (n 80) paras 51-52.

191 Lerner (n 132) 49; Buyse (n 56) 792.

192 John Dovido and James Jones, 'Prejudice, Stereotyping, and Discrimination' in Eli Finkel and Roy Baumeister (eds), *Advanced Social Psychology* (2nd edn, Oxford University Press 2019) 279.

193 *Mohamed Rabbae, ABS and NA v The Netherlands* (n 74) para 7.

194 UN Human Rights Committee, 'General Comment No 18: Non-discrimination' (10 November 1989) UN Doc HRI/GEN/1/Rev.9 (Vol. I) para 13.

195 Schabas (n 81) 584.

196 Notably, the ECtHR stated that 'merely inciting a difference in treatment [do] not necessarily amount to inciting discrimination, in *Baldassi and Others v France* App No 15271/16 and 6 others (ECtHR, 11 June 2020) para 64.

introduced varying proposals.[197] The Human Rights Committee and CERD, however, have rarely specified which prohibited harm applies in their decision – discrimination, hostility or violence – and have not clarified the threshold-question. In *Mohammed Rabbae A.B.S and N.A v Netherlands*, the Committee merely concluded that the expressions created 'discriminatory social attitude against the group and against them as members of the group,'[198] and in the *Jewish Community of Oslo et al. v Norway*, the CERD found statements containing discriminatory falsehoods to constitute incitement at least to racial discrimination, if not violence,'[199] because they 'were of an exceptionally/manifestly offense character.'[200]

Notwithstanding valid concern fearing the overbreadth of the provisions, including incitement to discrimination serves a clear and legitimate purpose by providing a flexible and nuanced framework to address complex societal harms. The risks of discriminatory propaganda were even among the reason to include incitement to discrimination in Article 20 in the first place. During the drafting of the ICCPR, members of the Commission on Human Rights stressed that 'in view of the experience with manipulative power of modern propaganda, the evil should be attacked at its roots', for which they considered far-reaching criminal prohibitions necessary.[201] The central position of equality throughout the ICCPR supports that 'to combat the roots of the main causes of their systematic violation (wars, as well as racial, national and religious discrimination) by way of preventive prohibitions in the area of formation of public opinion.'[202] This historical rationale proves especially relevant to contemporary discriminatory disinformation.

As Evatt and Kretzmer have noted in their concurring opinion in *Faurisson v. France*, freedom 'from discrimination on grounds of race, religion and national origins' extends to 'incitement to such discrimination.'[203] In 'particular social and historical context', they argue, 'statements that do not meet the strict legal criteria of incitement can be shown to constitute part of a *pattern* of incitement against a given racial, religious or national group', which must fall within the provision's scope. Especially since 'those interested in spreading

197 Temperman (n 60) 189-190; Nazila Ghanea, 'Expression and Hate Speech in the ICCPR: Compatible or Clashing' (2005) 5 Religion & Human Rights, 171-190, 175.

198 *Mohamed Rabbae, ABS and NA v The Netherlands* (n 74) para 9.6

199 ICERD, *The Jewish Community of Oslo and Others v Norway*, Communication No 30/2003 (15 August 2005) UN Doc ICERD/C/67/D/30/2003, paras 2.1, 10.4.

200 *Ibid.* paras 2.1, 10.5 (The speech made during a commemoration march for Nazi leader Rudolf Hess, claiming that 'our people and country are being plundered and destroyed by Jews, who suck our country empty of wealth and replaced it with immoral and un-Norwegian thoughts', calling upon the group to 'follow in […] the footsteps' or Hess and Hitler, beyond making extensive unfounded accusations of the robbing, raping and killing of Norwegian by immigrants).

201 Schabas (n 81) 578.

202 *Ibid.* 577; Partsch (n 127) 277.

203 *Faurisson v. France* (n 136) para 4.

hostility and hatred adopt sophisticated forms of speech that are not punishable under the law against racial incitement, even though their effect may be as pernicious as explicit incitement, if not more so'.[204] ICERD, without expressly clarifying the scope, recognised this dynamic by prohibiting 'propaganda activities, which promote and incite racial discrimination' – propaganda being described by Thornberry as 'appealing to the emotions rather than reason and is intended to persuade to a point of view.'[205] By explicitly mandating that propaganda activities that incite discrimination shall be prohibited, the Convention recognises that manipulative information campaigns can rise to this level. This enables direct application of these norms to contemporary forms of discriminatory disinformation. The prohibition of incitement to discrimination under both Conventions thus provides a crucial legal framework for addressing discriminatory disinformation, which operates through precisely the manipulative mechanisms these provisions were designed to counter.

5.4.5.2 Incitement to Hostility

Unlike discrimination and violence, 'hostility' is neither included nor defined in international human rights law, primarily because it is not a legally definable concept.[206] While discrimination and violence describe action – often motivated by mental states – hostility connotates a mental state in itself.[207] This explains why 'hostility' is regularly equated with 'hatred',[208] despite the logical problems this creates regarding the qualified nature of Article 20(2) ICCPR. For there to be a qualification of 'incitement to hatred', the audience would merely have 'to copy the inciter's 'hatred'.[209] Such interchangeability would establish an illogically low threshold ignoring the provision's express distinction between 'advocacy for hatred' and 'incitement to hostility.'

Both the ICCPR and ICERD, however, are not concerned with hostile thoughts,[210] but with 'tangible repercussions of the poisoned mind.'[211] In

204 *Faurisson v. France* (n 136) Individual Opinion of Evatt and Kretzmer, para 4.
205 Thornberry (n 92) 286.
206 McGonagle (n 7) 273 citing Partsch (n 127) 228.
207 JK Miles, 'Hatred, Hostility and Defamation' (2011) 25 International Journal of Applied Philosophy 1, 28.
208 Camden Principles (n 100) 10; UN Human Rights Council, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Addendum' (n 145) fn 5 ('[n]ational systems should make clear that […] that the terms "hatred" *and* "hostility" refer to intense and irrational emotions of opprobrium, enmity and detestation towards the target group' [emphasis added]).
209 Temperman (n 60) 188.
210 Toby Mendel, 'Study on International Standards Relating to Incitement to Genocide or Racial Hatred' (UN Special Advisor on the Prevention of Genocide 2006) 28.
211 Temperman (n 60) 188.

light of the absolute nature of freedom of thought,[212] incitement to *hostility* as a mental state should be understood as incitement to *hostile acts*, at least in the context of Article 20(2) ICCPR. Attempts to differentiate between 'hatred' and 'hostility' lead to a similar outcome. Nowak's Commentary and the UN Special Rapporteur on Freedom of Expression – while recognising their conceptual adjacency – define hostility as 'a *manifestation of hatred* beyond of mere State of mind (emphasis added),'[213] which takes the form of actual harmful effects.[214] While hatred has a strong internal and subjective element, hostility suggests an 'attitude displayed externally'.[215] Notwithstanding the solidity of this interpretation, both treaties inevitably address mental processes to some extent. For example, in *Ross v Canada*, the Human rights Committee acknowledged that 'the principles reflected in Article 20(2)' entails that 'restrictions [on freedom of expression] may be premised on statements which are of a nature as to raise or strengthen Anti-Semitic *feeling*, in order to uphold the Jewish communities' right to be protected from religious hatred (emphasis added)'.[216]

The line between hostility and the other proscribed harms remains poorly defined, and there does not appear to be a need to clarify this; neither the Human Rights Committee nor commentators seem to take on the matter.[217] Temperman nevertheless suggests that the difference between hostile acts and violence 'appears to be but a matter of scale,'[218] while others view discrimination, hostility and violence as conceptually distinct phenomena, which cover fundamentally different mental processes and behaviour.[219] The textual

212 General Comment No 34 (n 80) para 5; Sjors Ligthart *et al.*, 'Rethinking the Right to Freedom of Thought: A Multidisciplinary Analysis' (2022) 22 Human Rights Law Review 4, 10; Patrick O'Callaghan et al, 'The Right to Freedom of Thought: An Interdisciplinary Analysis of the UN Special Rapporteur's Freedom of Thought' (2024) 1 The International Journal of Human Rights 1-23; UN Special Rapporteur on Freedom of Religion or Beliefs, 'Interim report of the Special Rapporteur on freedom of religion or belief, Ahmed Shaheed' (5 October 2021) UN Doc A/76/380 para 4; UN Human Rights Committee, 'General Comment No. 22: Article 18 (Freedom of Thought, Conscience or Religion)' (30 July 1993) UN Doc CCPR/C/21/Rev.1/Add.4 paras 1 and 3.

213 Schabas (n 81) 587; UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression' (n 17) para 44; ARTICLE19 (n 112) 19 ('"[h]ostility" shall be understood as a manifested action of an extreme state of mind. Although the term implies a state of mind, an action is required').

214 UNGA 'Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Frank La Rue' (n 17) para 45(e)

215 McGonagle (n 7) 273 citing Partsch (n 127) 228 ([t]he 'strong connotation of war' attached to 'hostility' due to its terminological centrality in international humanitarian law' is also not addressed in IHLR).

216 *Malcolm Ross v. Canada* (n 103) para 11.5.

217 Temperman (n 60) 189; Draft General Comment No 34 (n 96) para 53 ('[i]t would be sufficient that the incitement relates to any of the three outcomes: discrimination, hostility or violence').

218 Temperman (n 60) 188.

219 Miles (n 207) 25-32; Dias (n 66) 403-404.

position of hostility between discrimination and violence suggests it exceeds discriminatory acts while falling short of physical violence – potentially encompassing psychological violence, threats and other aggressive behaviour. Focussing on hostility as a mental state instead would frame incitement to hostility as creating a hostile environment of 'poisoned minds' as a precursor for violence, where enmity and antagonism towards a particular groups become dominant sentiments.[220]

Discriminatory disinformation plays a central role in creating hostile environments, as illustrated in the *Ross v. Canada* case: antisemitism represents the most striking example, being intrinsically a form of discriminatory disinformation embedded in many hostile extremist ideologies.[221] The use of false and misleading information to incite may target all groups protected under hate speech prohibitions.[222] While incitement to discrimination specifically disadvantages individuals or groups, incitement to hostility through disinformation creates a broader climate of hatred that fundamentally undermines a group's societal standing. Discriminatory disinformation campaigns systematically employ false narratives to delegitimise targeted communities, creating environments where exclusionary attitudes become normalised. This distinction – though not incontestable – facilitates understanding how the different categories of proscribed harm correspond to distinct rhetorical strategies and psychological mechanisms. Accordingly, discriminatory disinformation operates most powerfully in the realm of cultivating hostile environments through sustained, coordinated manipulation of public perception, and thus can be prohibited as such.

5.4.5.3 Ideas of Racial Superiority

Article 4 ICERD extends beyond the ICCPR framework by mandating criminal measures for 'all dissemination of ideas based on racial superiority or hatred.' This provision significantly broadens the scope of prohibited expressions and was initially added to the 'lexicon of prohibited acts' to ensure that 'incitement and violence are not the only forms of activity to be sanctioned.'[223] While the inclusion reflects a recognition of the danger of racist discourse in stirring up hatred and violence, it lacks explicit requirements of intent or demonstrable

220 Lerner (n 132) 52; Wibke Timmermann, 'The Relationship Between Hate Propaganda and Incitement to Genocide: A New Trend in International Law Towards Criminalization of Hate Propaganda' (2005) 18 Leiden Journal of International Law 257-282, 264.

221 Radicalisation Awareness Network, 'Antisemitism as a Part of Almost All Extremist Ideologies and Narratives' (29-30 March 2022) Accessed 8 September 2024.

222 Section 5.2.1.2 'Object of the Message'.

223 Thornberry (n 92) 278.

harm.[224] This diverges from the ICCPR's approach – a 'striking contrast to the leading conventions and other non-treaty reference points in international law.'[225]

The expansive formulation has created implementation challenges. Despite initial concerns about incompatibility with freedom of expression, the ICERD Special Rapporteur in 1983 affirmed that 'the mere act of dissemination is penalised, despite lack of intention to commit an offence and irrespective of the consequences of the dissemination, whether it be grave or insignificant.'[226] This creates *de facto* strict liability – a position that, as Thornberry observes, 'does violence to basic principles of criminal liability in many if not most jurisdictions.'[227] Consequently, States have been hesitant to fully subscribe to and implement the prohibition, resulting in numerous reservations.[228]

General Recommendation 35 (2013), however, marked a significant shift, narrowing the scope of the provision by stipulating that criminalisation 'should be reserved for serious cases to be proven beyond reasonable doubt, while less serious cases should be addressed by means other than criminal law, taking into account, *inter alia*, the nature of the impact on targets persons and groups.'[229] While this interpretation moves closer to the ICCPR,[230] it remains curious that out of the five categories of offences in the General Recommendation, the 'dissemination of ideas based on racial or ethnic superiority or hatred, by whatever means' is the only one without an explicit requirement that such speech must amount to incitement.[231]

While ideas of racial superiority are always characterised by some form of disinformation, the Recommendation specifically addresses the (un)lawfulness of public denials or attempts to justify crimes of genocide and crimes against humanity.[232] These are clear illustrations of discriminatory disinformation,[233] instrumentalising racial superiority narratives and falsehoods about group identity. Prohibiting these types of expression, the ICERD notes, is 'a

224 Tarlach McGonagle, 'General Recommendation 35 on Combatting Racist Hate Speech' in David Keane and Annapurna Waughray (eds), *Fifty years of the International Convention on the Elimination of All Forms of Racial Discrimination* (Manchester University Press 2017) 252.
225 McGonagle (n 7) 284; Thornberry (n 92); Aswad and Kaye (n 142) 177-179.
226 Jose D Inglees, 'Study on the Implementation of Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination' (18 May 1983) UN Doc A/CONF.119/10, para 83.
227 Thornberry (n 92) 292.
228 Clooney and Gardoll (n 75) 177; Thornberry (n 92) 278-280; Temperman (n 60) 221-222.
229 ICERD General Recommendation 35 (n 89) para 12.
230 Clooney and Gardoll (n 75); McGonagle (n 241) 255-256; Temperman (n 60) 223; Aswad and Kaye (n 142) 177-179; Thornberry (n 92) 181.
231 ICERD General Recommendation 35 (n 89) para 13.
232 Thornberry (n 92) 285.
233 Sections 5.2.1.2 'Object of the Message'; 5.2.1.3 'Objective of the Author'.

forthright expression of the preventive function of the Convention and [is] an important complement to the provisions on incitement.'[234]

Within discriminatory disinformation and racial discrimination, the dimension of scientific disinformation represents a particularly insidious manifestation. Illustrated by Lerner, 'it should not be forgotten […] that in the past many books and papers aimed at disseminating racial hatred adopted the external form of "scientific" books or studies. The Nazi regime was especially prolific in the production of such studies.'[235] In recent years, new vectors have emerged for old narratives. Predatory and pseudo-journals and publishers, alongside occasional false or misleading publications by recognised scholars,[236] are lending perceived credibility to questionable ideas and racial superiority theories on an unprecedented scale.[237] This creates sophisticated and coordinated channels for discriminatory disinformation that engages Article 4(b) obligations for States to prohibit organisations and 'other propaganda activities', which promote and incite discrimination under the Convention. This dimension offers a valuable framework for addressing discriminatory disinformation promoting racial superiority ideas, capturing content that might evade traditional incitement standards. However, its breadth poses risks to legitimate speech and potentially academic freedom, and the absence of jurisprudential or other guidance available leaves actual implementation uncertain.

5.4.6 Threats or Incitement to Violence

The third proscribed harm in Article 20(2) ICCPR initially appears straightforward and familiar to international law compared to discrimination and hostility. Violence, Nowak's Commentary suggests, does not require any special definition,[238] and Lerner agrees that punishing acts of violence and incitement to such carries no particular difficulties.[239] A closer examination, however, reveals subtle ambiguities, particularly in the text of Article 4 ICERD, which refers to 'all acts of violence or incitement to such acts', while its accompanying General Recommendation addresses 'threats or incitement to violence'.[240]

234 ICERD General Recommendation 35 (n 89) para 11.
235 Lerner (n 132) 49.
236 The New York Times, 'Harvard Scholar Who Studies Honesty Is Accused of Fabricating Findings' The New York Times (25 June 2023) Accessed 23 January 2025.
237 Brian G Southwell *et al.*, 'Defining and Measuring Scientific Misinformation' (2022) 700 The ANNALS of the American Academy of Political and Social Science 1, 9; Jevin D West, 'Misinformation In and About Science' (2021) 118 PNAS 15, 2-5.
238 Schabas (n 81) 587.
239 Lerner (n 132) 49.
240 ICERD General Recommendation 35 (n 89) para 3.

The WHO World Report on Violence and Health from 2002 provides an often-cited definition of violence as 'the intentional use of physical force or power against another person, or against a group or community that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment, or deprivation.'[241] This definition expands violence beyond mere physical acts to include psychological and potentially structural forms of violence without immediate physical manifestations. These, and comparable understandings, have generated divergent reactions. On one hand, experts suggest moving away from narrow interpretations of violence as merely physical acts, to better perform the preventive function of the Convention.[242] In contrast, others argue for a more restrictive interpretation – or even limitation – in both Conventions to physical violence as the singular, at least most prominent, proscribed harm.[243] Either way, certain forms of incitement to physical violence remain unequivocal, direct ('[t]he Jews in Russia must be killed. They must be exterminated root and branch'[244]) or indirect ('go to work' meaning to '[g]o kill the Tutsis and Hutu political opponents of the interim government').[245] International courts (5.5) have extensively explored these forms of incitement.

Further inclusion of psychological and even structural violence, however, challenges traditional conceptions and blurs the lines between hostility and violence. While various factors could distinguish between these concepts – including scale, intensity or cruelty[246] – a more binary approach separating physical and non-physical or mental harm – offers greater clarity.[247] Accepting that incitement to violence includes impairment of a person's psychological integrity has direct implications for discriminatory disinformation.[248] Even if exposure to discriminatory disinformation itself does not constitute psychological violence, its proliferation has incited such harm directly or indirectly. A striking example is the sustained practice of LGBTQI+ conversion therapy,

241 ARTICLE19 (n 112) 19.

242 Gelber (n 6); Fino (n 185) 202; Alexander Tsesis, 'Inflammatory Speech: Offense Versus Incitement' (2013) 97 Minnesota Law Review 1145-1196; Susan Benesch, 'Countering Dangerous Speech: New Ideas for Genocide Prevention' (11 February 2014) Accessed 9 September 2024.

243 Temperman (n 60) 189; ARTICLE 19 (n 112) 19; UN Strategy and Plan (n 8) 3-4.

244 Thomas J Dodd, *Prosecution oral presentation against Fritzsche* (University of Connecticut Archives & Special Collections at the Thomas J Dodd Research Centre) Accessed 8 September 2024.

245 *Prosecutor v Ruggiu* (Judgment and Sentence) ICTR-97-32-I (1 June 2000) para 44(iii).

246 Temperman (n 60) 188; Rabat Plan of Action (n 81) para 29.

247 Jan Christoph Bublitz and Reinhard Merkel, 'Crimes Against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination' (2024) 8 Criminal Law and Philosophy 51-77, 58.

248 This qualification of psychological violence is loosely based on Article 33 of the Council of Europe Convention on Preventing and Combatting Violence against Women and Domestic Violence (adopted 11 May 2011, entered into force 1 August 2014) CETS No 210.

with disinformation about its effectiveness facilitating and fuelling practices that cause profound psychological damage.[249]

Finally, discriminatory disinformation also contributes to the 'positive feedback loop' between online hate speech and offline violence – patterns where violence leads to more messages advocating violence and vice-versa.'[250] Narratives fabricating non-existing threats of imminent risks of violence or that falsely attribute the use of violence to a particular group in society strengthen this cycle. This demonstrates a complex interaction between physical and psychological violence, structural incitement and manipulation. Developments that have thus far been insufficiently addressed by the doctrine on incitement to violence.

5.4.7 Interim Conclusion

This analysis demonstrates that international hate speech prohibitions under Article 20(2) ICCPR and Article 4 ICERD provide a viable legal framework for addressing discriminatory disinformation, despite certain doctrinal uncertainties. The systematic examination of the provisions' individual components revealed the needed convergence. First, evolving interpretations of the protected group coverage enables application to almost all discriminatory disinformation, regardless of whether it targets groups on racial or ethnic grounds or employs, *inter alia*, gender-disinformation. Second, while disinformation inherently involves an intent to cause harm, it may not always centre on promoting hatred as its primary motivation. Unless discriminatory disinformation has an exclusive commercial incentive, it likely satisfies the intent requirements for advocacy under Article 20(2) ICCPR. In addition, while discriminatory disinformation may not always explicitly advocate hatred, it consistently contributes to environments where hateful attitudes become normalised – precisely the condition these prohibitions aim to prevent.

Where disinformation mostly does not directly incite violence, the contextual approach to incitement in relation to the proscribed non-physical harms is particularly suited to discriminatory disinformation's characteristics. The frameworks' emphasis on contextual factors in determining whether expression amounts to incitement, including content, form, and presentation of information, as well as the economic, social, and political climate, enables a flexible structure. It accounts for the consideration of cumulative cognitive and behavioural influence, the information environment and contextual assessment of 'likelihood of harm'. The analysis of incitement to non-physical harms –

249 United Nations Press Release, 'One UN Human Rights Expert's Fight to Eliminate 'Conversion Therapies'' (18 February 2022).

250 Alexandra Olteanu *et al.*, 'The Effect of Extremist Violence on Hateful Speech Online' (2018) 12 Proceedings of the International AAAI Conference and Web and Social Media 1.

discrimination, hostility, and the dissemination of ideas based on racial superiority – provides the most direct legal pathway for addressing discriminatory disinformation, capturing how false or misleading information systematically targets vulnerable groups and undermines social cohesion. The identified 'positive feedback loop' between online disinformation and offline violence further evidences the centrality of discriminatory disinformation in prohibited incitement to physical harm – either direct or indirect. While interpretative challenges remain – especially regarding intent thresholds and proximity standards – the fundamental prohibitions against advocacy of hatred that constitutes incitement to discrimination, hostility, or violence demonstrate sufficient flexibility to encompass the threat of emerging disinformation.

5.5 Disinformation and Incitement, Denial and False Allegations of Genocide

> *'All forms of public propaganda tending by their systematic and hateful character to promote genocide, or tending to make it appear as a necessary, legitimate or excusable act shall be punished.'*
>
> – 1947 Draft Convention on the Crime of Genocide

Hate propaganda is a widely recognised *condicio sine qua non* for creating conditions conducive to genocide.[251] Emerging evidence demonstrates disinformation's role in enabling, furthering and facilitating international crimes, exemplified by ISIS and the Yazidi minority, the Rohingya in Myanmar and the Uyghur repression in Xinjiang, China. Concurrently, false allegations of genocide have been used by Russia as a pretext for its large-scale invasion of Ukraine in 2022, while the denial of mass atrocities, most notably the Holocaust, undermines social cohesion and stability, nationally and internationally. The legal concept of genocide suffers from inflation through persistent misuse in political and societal discourse, and is increasingly instrumentalised by propagandists, authoritarian regimes and media sensationalism.[252] A development exacerbated by disinformation.

This part of chapter five examines legal frameworks applicable to these three dimensions of genocide-related disinformation: as a form of and/or as

251 Paul R Bartrop, *Genocide and Propaganda: A Primary Source Collection* (Bloomsbury Academic 2025); Leo Kuper, *Genocide: Its Political Uses in the Twentieth Century* (Yale University Press 1981); Staub (n 51); Gregory Stanton, *The Eight Stages of Genocide* (Genocide Watch 1996) Accessed 14 April 2024.

252 Eric A Heinze, 'The Rhetoric of Genocide in U.S. Foreign Policy: Rwanda and Darfur Compared' (2007) 122 Political Science Quarterly 3, 359-383; Oksana Dudko, 'A Conceptual Limbo of Genocide: Russian Rhetoric, Mass Atrocities in Ukraine, and the Current Definition's Limits' (2022) 64 Canadian Slavonic Papers 2-3, 133-140.

a tool for incitement to commit genocide, genocide denial, and false allegations of genocide. Each dimension illustrates how disinformation is instrumentalised to manipulate perceptions, incite violence, evade responsibility and distort history to the detriment of victims. While certain aspects fall under international criminal law, others are primarily subject to the human rights framework (5.4).

Following the structure of preceding chapters, the analysis takes the legal concept of 'genocide' as its point of departure (5.5.1). It then examines international law's treatment of discriminatory disinformation as incitement to genocide – an inchoate offence firmly established in international criminal statutes and jurisprudence (5.5.2). Beyond this recognised legal framing, the convergence between disinformation and genocide denial as a tool of incitement to genocide and prohibited hate speech is explored (5.5.3), followed by the legal limitations on disinformation constitutive of false genocide allegations, often as a pretext for violence (5.5.4).

5.5.1 Scope of the Crime of Genocide

Raphael Lemkin formulated genocide as the 'crime of crimes' in a time when the unprecedented role of propaganda in the commission of atrocities was revealed.[253] As a core concern of the international community as a whole, the prohibition to commit genocide and the obligations to punish and prevent its occurrence were codified, with violations considered internationally wrongful acts as well as grounds for individual criminal responsibility.

As customary international law and recognised *ius cogens*,[254] genocide is defined in Article II of the Convention on the Prevention and Punishment of the Crime of Genocide (Genocide Convention). It encompasses the commission of punishable acts, including 'killing of members of a group' or causing serious bodily or mental harm to members of the group' accompanied by the 'intent to destroy, in whole or in part, a national, ethnical, racial or religious group, as such […].'[255] Extensive scholarship demonstrates that genocide comprises genocidal *acts* against the physical or psychological integrity of

253 Raphael Lemkin, *Genocide as a Crime Under International Law* (1948) UN Bull 4:70-71, 70.

254 *Reservations to the Convention on the Prevention and Punishment of the Crime of Genocide* (Advisory Opinion) [1951] ICJ Rep 15, 23; *Application of the Convention on the Prevention and Punishment of the Crime of Genocide* (Bosnia and Herzegovina v Serbia and Montenegro) (Judgment) [2007] ICJ Rep para 161; *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T, T Ch I (2 September 1998) para 295; *Prosecutor v Krstic* (Judgment) IT-98-33-T (2 August 2001) para 541.

255 Convention on the Prevention and Punishment of the Crime of Genocide (adopted 9 December 1948, entered into force 12 January 1951) 78 UNTS 277 [hereafter: Genocide Convention].

members of a specific group, yet may also occur through omission.[256] In both scenarios, the specific 'intent to destroy' constitutes the crime's essence,[257] distinguishing it from other international crimes.

While the Convention primarily addresses physical destruction, the concept of 'cultural genocide' – the systematic erasure of a group's cultural and identity – was considered during the drafting process but ultimately excluded from the final text.[258] Despite scholarly recognition that cultural destruction often represents 'an intrinsic characteristic of every process of genocide,'[259] cultural genocide is not a prohibited crime under international law *per se*.[260] Though the present focus is limited to established legal frameworks, disinformation as a tool of cultural genocide warrants consideration in future debates on this concept and its regulation. In brief, cultural genocide encompasses destruction of a community's distinctive spiritual, material, intellectual, and emotional features, often manifesting through various disinformation strategies: denigrating cultural practices, dismissing indigenous languages as 'useless' and 'primitive' – facilitating their abandonment, appropriating cultural elements, manipulating educational and artistic cultural content and engaging with historical revisionism.[261] Such perpetual denial, denigration or manipulation of a group's cultural features and practices frequently appear predominantly alongside – or as an integral part of – ongoing or imminent physical genocides.[262] Contemporary examples – such as Russia's campaigns against

256 William A Schabas, *Genocide in International Law: The Crime of Crimes* (2nd edn, Cambridge University Press 2009); Ralph Henham and Paul Behrens (eds), *The Criminal Law of Genocide* (Ashgate 2007); John Quigley, *The Genocide Convention: An International Law Analysis* (Ashgate 2006); Onur Uraz, *Classifying Genocide in International Law: The Substantiality Requirement* (Routledge 2023); Florian Jeßberger, 'The Definition and the Elements of the Crime of Genocide' in Paola Gaeta (ed), *The UN Genocide Convention: A Commentary* (Oxford University Press 2009) 87-90.

257 Jeßberger (n 256) 93.

258 Alisa Novic, *The Concept of Cultural Genocide: An International Law Perspective* (Oxford University Press 2016) 4; Martin Hamilton, 'The Concept of Cultural Genocide' in Claire Finkelstein (ed), *The Preservation of Art and Culture in Times of War* (Oxford University Press 2022) 140-151.

259 *Ibid.* 4, 8.

260 Daphne Anayiotos, 'The Cultural Genocide Debate: Should the UN Genocide Convention Include a Provision on Cultural Genocide, or Should the Phenomenon be Encompassed in a Separate International Treaty' (2009) 22 New York International Law Review 2; Yvonne Donders, 'Old Cultures Never Die? Cultural Genocide in International Law' in Ineke Boerefijn *et al.* (eds), Human Rights and Conflict, Essays in Honour of Bas de Gaay Fortman (Intersentia 2012); David Nersessian, 'A Modern Perspective: The Current Status of Cultural Genocide Under International Law' in Jeffrey S Bachman (ed), *Cultural Genocide: Law, Politics, and Global Manifestations* (Routledge 2019).

261 Suzanne Romaine, 'The Global Extinction of Language and Its Consequences for Cultural Diversity' in Heiko Maarten *et al.* (eds), *Cultural and Linguistic Minorities in the Russian Federation and the European Union* (Springer 2015) 35-36; Tove Skutnabb-Kangas, *Linguistic Genocide in Education – or Worldwide Diversity and Human Rights* (Routledge 2000),

262 On cultural genocide and propaganda during WOII, see Anayiotos (n 260) 104.

Ukraine's cultural identity,[263] and Iran's systematic erasure of the Baha'i community –[264] illustrate disinformation as *de facto* cultural genocide, yet these campaigns generally fall outside genocide's legal definition under current international law.

From this foundation, the legal analysis focuses not on genocide's *actus reus* and *mens* rea, but on the *mens rea* of incitement, denial and false allegations of genocide to disinformation. Most experts interpret 'intent to destroy' through a purpose-based approach,[265] meaning that 'the perpetrator acts with the aim, purpose or desire to destroy.'[266] In contrast, a minority view favours a knowledge-based approach, suggesting that knowledge of the destruction should be sufficient.[267] These approaches are framed in terms of direct or indirect intent, imposing different thresholds in terms of intensity and standards. They imply a dual intent requirement in disinformation as incitement to genocide: the underlying intent to destroy, and the intent to incite. For genocide denial and false allegations of genocide, this dual-intent requirement is only relevant when aiming to prove that these forms of disinformation rise to the level of incitement to genocide.[268]

5.5.2 Incitement to Genocide

Without diminishing the responsibility of direct perpetrators, a historical review of the gravest crimes in 20th century reveals how inciting propaganda

263 Jade McGlynn, 'Russia Is Committing Cultural Genocide in Ukraine' (23 April 2024) Foreign Policy, Accessed 9 October 2024; New Lines Institute for Strategy and Policy and Raoul Wallenberg Centre for Human Rights, 'An Independent Legal Analysis of the Russian Federation's Breaches of the Genocide Convention in Ukraine and the Duty to Prevent'(May 2022) 1-2, 13-20; Ian Garner, '"We've Got to Kill Them": Responses to Bucha on Russian Social Media Groups' (2023) 25 Journal of Genocide Studies 418, 418-425; Michael Schwirtz, Maria Varenikova and Rick Gladstone, 'Putin Calls Ukrainian Statehood a Fiction. History Suggests Otherwise' The Washington Post (21 February 2022); Max Fischer, 'Word by Word and Between the Lines: A Close Look at Putin's Speech' The Washington Post (23 February 2022); Peter Dickinson, 'Putin's New Ukraine Essay Reveals Imperial Ambitions' (Atlantic Council, 15 July 2021).

264 Moojan Momen, 'The Baha'I community of Iran: cultural genocide and resilience' in Jeffrey S Bachman (ed), *Cultural Genocide: Law, Politics, and Global Manifestations* (Routledge 2019) 250-259 ('black propaganda and disinformation' have been 'fundamental to both the physical and cultural genocide of Baha', portraying them as 'a foreign creation, designed to weaken Iran', with an 'artificial religion created by British' and 'linked to foreign powers' plotting to destroy Islam').

265 Sangkul Kim, *A Collective Theory of Genocidal Intent* (Asser Press 2016) 18-19.

266 Jeßberger (n 256) 105.

267 *Ibid.* 105; Kim (n 265) 22-30.

268 Jeßberger (n 256) 87; Kim (n 265) 5; Even if knowledge cannot replace the specific intent, knowledge of participating in an extermination can indicate the presence of the required intent, in Gerhard Werle and Florian Jeßberger, *Principles of International Criminal Law* (3rd edn, Oxford University Press 2014) 314.

facilitates, amplifies and sustains atrocities.[269] Genocide does not occur spontaneously, and those who organise it are not necessarily the only perpetrators.[270] As Applebaum noted regarding the war in Ukraine, 'while not every use of genocidal hate speech leads to genocide, all genocides have been preceded by genocidal hate speech.'[271] Since Julius Streicher's indictment on 8 October 1945,[272] the significance of propaganda and information operations has only expanded, challenging courts and tribunals to address their harmful consequences. While the jurisprudence of the Nuremberg Tribunals, the ICTY and the ICTR suggests that propaganda inciting or contributing to genocidal violence can be prohibited as direct incitement to commit genocide (as a substantive crime) or as a mode of liability, it remains an 'unsettled area of international criminal law.'[273] Existing case has thus far not treated propaganda or disinformation *per se* as incitement to genocide.

The prohibition of incitement to commit genocide has crystallised into customary international law,[274] is codified in various international instruments and consistently applied by courts and tribunals. Article III(c) of the Genocide Convention authoritatively states that 'direct and public incitement to commit genocide' […] 'shall be punishable'.[275] Similar wording appears in Article 3(c) of the Statute of the International Tribunal for Rwanda (ICTR) and Statute of the International Criminal Tribunal for the Former Yugoslavia (ICTY).[276] A verbatim formulation is included in Article 25(2)(e) of the Rome Statute, whereby – though situated amongst grounds of liability of the other core crimes – the Statute expressly limits criminal responsibility for incitement as a substantive crime to the crime of genocide – despite repeated efforts to

269 Timmerman (n 50) 3; Tonja Salomon, 'Freedom of Speech v Hate Speech: The Jurisdiction of 'Direct and Public Incitement to Commit Genocide' in Ralph Henham and Paul Behrens (eds), *The Criminal Law of Genocide* (Ashgate 2007) 141-142.

270 Hefti and Ausserlandscheider (n 174) 30; Elise van Sliedregt, 'Criminalizing of Crimes against Humanity under National Law' (2018) 16 Journal of International Criminal Justice 729-749, 735 ('promoting an individual to commit a crime may be even more reprehensible than assisting someone who has already decided to commit a crime').

271 Anne Applebaum, 'Ukraine and The Words That Lead to Mass Murder' The Atlantic (25 April 2022).

272 Margarret Eastwood, *The Nuremberg Trial of Julius Streicher: The Crime of "Incitement to Genocide"* (Edwin Mellen 2011); *Streicher Judgement* (1946) 22 Trial of German Major War Criminal 501.

273 Richard Ashby Wilson and Metthew Gillett, 'The Hartford Guidelines on Speech Crime in International Criminal Law' (2020) Faculty Articles and Papers, 5 [hereafter: Hartfort Guidelines].

274 Jérôme de Hemptinne, 'Incitement' in Jérôme de Hemptinne *et al.* (eds), *Modes of Liability in International Criminal Law* (Cambridge University Press 2019) 403 ft. 115, 116 and 117.

275 Genocide Convention (n 255) Article 1.

276 UNSC, 'Statute of the International Criminal Tribunal for Rwanda' (adopted 8 November 1994, last amended 13 October 2006) UN Doc S/RES/955; UNSC, 'Statute of the International Criminal Tribunal for the Former Yugoslavia' (adopted 25 May 1993) UN Doc S/RES/827.

expand the crime of incitement to the other core crimes during the drafting conference.[277]

All these instruments pose 'rigid and demanding' thresholds for incitement to genocide.[278] Therefore, the number of convictions remain few, with significant acquittals before courts and tribunals. No State has ever been held legally responsible for incitement to commit genocide. Notwithstanding, the ICJ in the *Case Concerning the Application of the Convention of the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro*) addressed the matter implicitly. While the Court concluded that it was not proven that the 'organs or persons (i.e. organs of the FRY or persons acting under its instructions or under the effective control] incited the commission of acts of genocide,'[279] rather than rejecting State attribution incitement to genocide all together, the Court relied on the insufficiency of the evidence and the *de facto* status as organs of the State of the involved actors.[280] In other words, if 'precise and incontrovertible evidence'[281] would be presented in the form of government-authorised speeches, official documents etc. which incited people to commit genocide, 'the Court would be hard pressed to deny State responsibility for incitement.'[282]

Propaganda and Incitement

Several landmark convictions and acquittals have shaped the jurisprudential relation between propaganda and incitement.[283] During the IMT trials, Julius Streicher, in his position as editor, was convicted for publishing genocidal articles in *Der Stürmer*, essentially constituting prosecution for incitement.[284]

---

277 Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3 (Rome Statute) Article 25(2)(e); Benesch (n 64) 509; Schabas (n 256) 325 referencing UN Doc A/CONF.183/C.1/WGGP/L.4, 3, UN Doc A/CONF.183/C.1/L.76/Add.3, 2, UN Doc PCNICC/1999/DP.4/Add.3, 3. The current negotiations on a Crimes against Humanity treaty, following the Draft articles on the Prevention and Punishment of Crimes against Humanity (2019) may significantly broaden the scope of available incitement-provisions.

278 Harmen van der Wilt, 'Between Hate Speech and Mass Murder: How to Recognise Incitement to Genocide' in Harmen van der Wilt *et al.* (eds), *The Genocide Convention: The Legacy of 60 Years* (Brill 2012) 42.

279 *Application of the Convention on the Prevention and Punishment of the Crime of Genocide* (Bosnia and Herzegovina v Serbia and Montenegro) (Judgment) [2007] ICJ Rep para 417.

280 Jens D Ohlin, 'State Responsibility for Conspiracy, Incitement, and Attempt to Commit Genocide' in Paola Gaete (ed), *The UN Genocide Convention – A Commentary* (Oxford University Press 2009) 374-375.

281 *Application of the Convention on the Prevention and Punishment of the Crime of Genocide* (n 296) para 417.

282 Ohlin (n 280) 374-376.

283 Gordon (n 1); Timmerman (n 50); Temperman (n 60).

284 Jens D Ohlin, 'Incitement and Conspiracy to Commit Genocide' in Paola Gaete (ed), *The UN Genocide Convention – A Commentary* (Oxford University Press 2009) 210; Diane F Orentlicher, 'Criminalizing Hate Speech In the Crucible of Trial: *Prosecutor v Nahimana*' (2006) 21 American University International Law Review 557, 582-584.

Nazi journalist Hans Fritzsche, though ultimately acquitted, was charged with inciting and encouraging the commission of war crimes 'by deliberately falsifying news to arouse in the German people those passions which led them to the commission of atrocities'.[285] Following the Rwandan genocide, multiple defendants were convicted for inciting statements on the radio, *Radio Télévision Libre des Mille Collines* (RTLM), in their capacity as radio moderator (Goerges Ruggiu), leadings editors (Jean-Bosco Barayagwiza and Ferninand Nahimana) and as the chief editor of an extremist newspaper (Hassan Ngeze).[286] Beyond this 'Media Trial', Simon Bikindi was found guilty of incitement to genocide through this song lyrics.[287] Cases before the ICTY, including *Prosecutor v. Vosjislav Šešelj, Prosecutor v. Radoslav Brdjanin, Prosecutor v. Milan Milutinović et al, Prosecutor v. Milomir Stakić* and *Prosecutor v. Dario Kordić & Mario Ćerkez*, further elucidate the speech-act complexities in the process of incitement.[288]

While these trials all subscribe to the possibility of propaganda being prohibited under international law when constituting incitement to genocide, doctrinally, it is not unambiguous. In the second half of the 20th century, scholars were critical of prohibiting forms of speech they deemed too remote – those lacking a direct call for immediate and concrete action.[289] For instance, Shaw noted in 1989 that Article III(c) of the Genocide Convention 'would not appear to be sufficiently broad to cover what may be termed public propaganda in favour of genocide,'[290] a position echoed by subsequent commentators.[291] This position aligns with the Ad Hoc Committee's rejection of an initial provision in the draft Genocide Convention that would have criminalised all forms of public propaganda tending to provoke genocide.[292] Nuancing

285 'Judgment of 1 October 1946' in *The Trial of German Major War Criminals: Proceedings of the International Military Tribunal sitting at Nuremberg, Germany*, Part 22 (22 August 1946 to 1 October 1946), 526.

286 *Prosecutor v Nahimana, Barayagwiza, and Ngeze* (Judgment and Sentence) ICTR-99-52-A (28 November 2007)

287 *Prosecutor v. Bikindi*, Case No. ICTR 01-72-T, Judgement para 4 (2 December 2008); Justin La Mort, 'The Soundtrack of Genocide: Using Incitement to Genocide in the Bikindi Trial to Protect Free Speech and Uphold the Promise of Never Again' (2009) 43 Interdisciplinary Journal of Human Rights Law 4; Heather MacLachlan, 'Music and Incitement to Violence: Anti-Muslim Hate Music in Burma/Myanmar' (2022) 66 Ethnomusicology 3, 410-442.

288 *Prosecutor v Vojislav Šešelj* (Judgment) IT-03-67-T (31 March 2016); *Prosecutor v Vojislav Šešelj* (Appeal Judgment) MICT-16-99-A (11 April 2018); *Prosecutor v Radoslav Brdjanin* (Judgment) IT-99-36-T (1 September 2004); *Prosecutor v Milan Milutinović et al.* (Judgment) IT-05-87-T (26 February 2009); *Prosecutor v Milomir Stakić* (Judgment) IT-97-24-T (31 July 2003); *Prosecutor v Dario Kordić and Mario Ćerkez* (Judgment) IT-95-14/2-T (26 February 2001).

289 Nehemiah Robinson, *The Genocide Convention* (New York: Institute for Jewish Affairs 1960) 67.

290 Malcolm N Shaw, 'Genocide in International Law' in Yoram Dinstein (ed), *International Law at a Time of Perplexity (Essays in Honour of Shabtai Rosenne)* (Martinus Nijhoff Dordrecht 1989) 811.

291 Timmerman (n 50) 212 citing Kai Ambos, *Der Ellgemeine Teil des Völkerstrafrecht: Ansätze einer Dogmatiserung* (2nd edn, Ducker and Humblot 2004) 415-416.

292 Robinson (n 289) 66; Ohlin (n 284) 212.

this denunciation, the Committee specifically rejected a provision criminalising '*indirect* propaganda' (emphasis added) – defined as 'propaganda which is intended to incite national, racial or religious hatred and to lead to genocide, but is not a direct incitement to genocide.'[293]

Disinformation permeates different stages of the genocide process, fuelling discriminatory attitudes, hostile environments and creating a fertile ground for ethnical violence. The scope of the crime of incitement to commit genocide, however, remains narrow, imposing high thresholds regarding the intent of perpetrator and the likelihood that the speech will trigger genocidal violence. Relevant provisions, commentaries and international jurisprudence have developed in fragmented, sometimes contradictory ways, complicating application to disinformation and resulting in a highly casuistic framework. Nevertheless, when disinformation contains expressions that are made publicly (5.5.2.1), directly 'incite' (5.5.2.2) and demonstrate a clear intent to incite to audience (5.5.2.3), it would be prohibited.

5.5.2.1 Public Nature

International instruments distinguish between public and private incitement, limiting the Genocide Convention's scope and subsequent statutes to public incitement. As previously noted,[294] the notion of 'public' is surrounded by definitional confusion,[295] leading to repeated calls for clarification,[296] particularly as some scholars opine that this ambiguity influenced acquittals during the ICTR trials.[297] It is nevertheless clear that the assessment depends on where the speech is uttered and who the audience is. While incitement – as a persuasive process – can evidently take place privately, such speech is excluded from the crime of incitement to genocide, yet it may constitute 'complicity in genocide' as a mode of liability.[298] The main consequence of this differentiation is that public incitement can be prosecuted even when genocide does not take place, whereas private incitement requires the underlying crime of genocide to have occurred.[299]

The meaning of 'public' has transformed in the online realm. In the 1990s, during the discussion on the Code of Crimes against the Peace and Security

293 UN Ad Hoc Committee on Genocide, 'Draft Report' (30 April 1948) UN Doc E/AC.25/W.1/Add.1, 3.
294 Section 5.4.2 'Advocacy and Intent'.
295 Gordon (n 1) 292; Schabas (n 256) 319.
296 Brendan Saslow, 'Public Enemy: The Public Element of Direct and Public Incitement To Commit Genocide' (2016) 48 Case Western Reserve Journal of International Law 1, 420.
297 *Ibid.* 420, 427.
298 Section 6.3.1 'Modes of Liability'.
299 Schabas (n 256) 31; International Law Commission, 'Report of the ILC on the work of its forty-eight session' (6 May – 26 July 1995) Official Record of the GA, 55th Session, Supplement No. 10, II YILC 2 1996, UN Doc A/51/10, 26-27.

of Mankind, the ILC noted that *public* incitement 'requires communicating the call for criminal action to a number of individuals in a public place or to members of the general public at large by means such as mass media.'[300] The ICTR Trial Chamber in *Prosecutor v. Akayesu* endorsed this approach, finding that the 'public requirement' is satisfied if speech addressed 'a number of individuals in a public place' or 'to members of the general public at large by such means as the mass media, for example, radio or television'.[301] While the number of addressees or the choice for a specific medium are not determinative,[302] these factors can substantiate a claim that the incitement was public.[303] The deciding element is 'that the appeal be aimed at a non-individualizable audience and thus create or enhance the danger of uncontrolled commission of the crime.'[304] In contrast, the ICTR Appeals Chamber characterised 'private incitement' as 'more subtle forms of communication such as conversations, private meetings, or messages,'[305] typically involving a select and limited audience.[306]

In contemporary information environments, the geographical dimension of the 'where' element has largely lost its relevance – 'public' comprises an international audience.[307] The Internet with its lack of geographical borders forms the primary vector for incitement. Adapting to this transformation, experts recognise that public incitement encompasses speeches and statements online,[308] the dissemination of (dis)information via open access Internet pages,

300 *Ibid.* 22.

301 *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T (2 September 1998) para 556; see also *Prosecutor v Ruggiu* (Judgment) ICTR-97-32-T (1 June 2000) para 17; *Prosecutor v Bikindi* (Decision on Defence Motion for Judgment of Acquittal) ICTR-2001-72-T (26 June 2007) para 29; *Prosecutor v Bagosora and others* (Decision on Motions for Judgement of Acquittal) ICTR-98-41-T (2 February 2005) para 22.

302 Ohlin (n 302) 186; De Hemptinne (n 274) 400, 403 citing *Prosecutor v Nzabonimana* (Appeals Chamber Judgment) ICTR-98-44D-A (29 September 2014) para 128; see also *Prosecutor v Kalimanzira* (Appeals Chamber Judgment) ICTR-05-88-A (20 October 2010) Separate Opinion of Judge Pocar, para 45.

303 Timmerman (n 50) 218 ft 137 citing *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T (2 September 1998) para 556; *Prosecutor v Ruggiu* (Judgment) ICTR-97-32-I (1 June 2000) para 17; *Prosecutor v Niyitegeka* (Judgment) ICTR-96-14-T (16 May 2003) para 431; *Prosecutor v Kajelijeli* (Judgment) ICTR-98-44A-T (1 December 2003) para 851; *Prosecutor v Nahimana and others* (Judgment) ICTR-99-52-T (3 December 2003) para 1011; *Prosecutor v Muvunyi* (Judgment) ICTR-2000-55A-T (12 September 2006) para 502.

304 Werle and Jeßberger (n 268) 323.

305 *Prosecutor v Kalimanzira* (Appeals Chamber Judgment) ICTR-05-88-A (20 October 2010) para 158.

306 Saslow (n 296) 430.

307 Robert Post, 'The Internet, Democracy and Misinformation' in András Koltay, Charles Garden and Ronald Krotoszynski (eds), *Disinformation, Misinformation and Democracy* (Cambridge University Press 2025) 47.

308 De Hemptinne (n 274) 398; Saslow (n 296) 419; Robert Cryer, 'Incitement' in Dinah Shelton (ed), *The Encyclopaedia of Genocide and Crimes against Humanity* (McMillan New York 2004) 252, 498; This coincides with the observations in section 4.4.3 'Incitement to Terrorism'.

and even emails addressed to large groups of individuals. Social media platforms – significant sources of hate speech and inflammatory language – balance between private and public channels. While their legal qualification of private, commercial, entities does determine this qualification, their infrastructure and accessibility of specific channels and spaces do. Platforms including X and YouTube would imply a public nature: these 'mass media' platforms are universally accessible with Internet connectivity as the sole limiting factor. Other platforms (e.g. Telegram and Discourse) have closed or semi-closed structures. Although they are public platforms, the structure of specific groups or channels targets a selected audience, suggesting a private nature. However, when these "private" groups include thousands of individuals, the qualification of a 'non-individualizable audience' becomes difficult to sustain.

The ICTR Appeals Chamber's distinction between public or private incitement based on the subtility of the communication creates more confusion than it solves. Applied to disinformation, while some forms involve subtle types of misdirection and manipulation at a private, individualised level, most campaigns expose large audiences to their message through the successor of 20th century 'mass media'. Private disinformation may have a more personalised and targeted effect on an individual's determination to commit a crime – being less "subtle"[309] – while public disinformation exploits the gradual effectiveness of subtle manipulation, being more effective in creating an atmosphere of violence and hatred as a prerequisite for genocidal violence to unfold.[310]

Consequently, platform structure, channel accessibility, the number of followers and/or people exposed to the information, should determine whether the dissemination of disinformation qualifies as public. The evolving nature of (new) online platforms requires that the notion is not interpreted statically. For instance, automated or fake personae entering small, closed online groups to strategically disseminate calls to engage in genocidal violence might strictly qualify as 'private communications', yet their coordinated employment across thousands of groups makes them functionally public to the intended audience. These patterns emerged in the inciting of genocide through disinformation against, *inter alia*, the Yazidi minority by ISIS and the persecution of the Rohingya population in Myanmar.

The 'public' element was intentionally included to limit the breadth of the prohibition of incitement to commit genocide.[311] However, the decreasing distinction between private and public spheres and audiences raises questions about this limitation's continued utility in the present form. This doubt

309 Wibke Timmerman, 'Incitement in International Criminal Law' (2006) 88 International Review of the Red Cross 864, 851.

310 *Ibid.*

311 Gordon (n 1) 188-189; De Hemptinne (n 274) 398; Saslow (n 296) 442.

strengthens proposals for eliminating the element of 'public' altogether.[312] Alternatively, the 'public' requirement could be adapted to the digital reality within existing frameworks. A doctrinal revision would need to consider 1) the algorithmic amplification capacity – content initially shared in private spaces becomes 'public' when platform algorithms substantially extend its reach beyond the original audience; 2) platform permeability – if messages in private spaces are designed to be easily shared across platform boundaries without access restrictions they may qualify as 'public'; and 3) coordinated deployment – (dis)information across multiple private channels becomes nevertheless public through synchronised distribution across multiple channels.

5.5.2.2 Direct Incitement

While incitement to commit genocide constitutes an elevated form of hate speech (5.4), its interpretative framework of 'incitement' largely coincides with that under Article 20(2) ICCPR and Article 4 ICERD, encompassing 'encouraging or persuading another to commit an offence'.[313] In the context of genocide, an explicit (textual) emphasis falls on the 'directness requirement' – the incitement must be potent enough to trigger criminal action, though actual provocation is not required.[314] The ICTR *Media case* authoritatively established that this 'potential of the communication to cause genocide' must be evaluated contextually.[315] As an inchoate offense,[316] incitement does not require causality between the speech and genocide.[317] Requiring otherwise would undermine the preventative nature of inchoate crimes and form 'a radical departure from at least a century of the criminal law of inchoate crimes'.[318]

Whether disinformation constitutes direct incitement depends on the context: is there a clear call to action? Existing factors for this evaluation that emerged in jurisprudence, however, misalign with disinformation's phased nature – harmful disinformation ultimately culminating in (genocidal) violence typically unfolds over time, involving multiple communication channels, and engaging various actors throughout its creation, production and dissemina-

312 Gordon (n 1) 294-295.
313 Andrew Ashworth, *Principles of Criminal Law* (4th edn, Oxford University Press 2003) 462.
314 Ohlin (n 284) 215-216.
315 *Prosecutor v Nahimana and others* (Judgment) ICTR-99-52-T (3 December 2003) para 1011.
316 Hartfort Guidelines (n 273) 41-43; Ashworth (n 313) 208; *Prosecutor v Nzabonimana* (Judgment) ICTR-98-44D-A (29 September 2014) para 234; *Prosecutor v Bikindi* (Judgment) ICTR-01-72-A (18 March 2010) para 149; *Prosecutor v Nahimana, Barayagwiza and Ngeze* (Judgment) ICTR-99-52-A (28 November 2007) 668, 720.
317 The Trial Chamber in *Prosecutor v. Akayesu* considered that it is not necessary that a causal connection between the inciting speech and the commission of genocide exists, though some authors doubt the appropriate practical interpretation, in Brabandere (n 68) para 23; Wilson (n 54) 25; Timmerman (n 50) 203; Schabas (n 256) 307; de Hemptinne (n 274) 398.
318 Wilson (n 54) 26.

tion.[319] A single piece of disinformation viewed in isolation is unlikely to rise to the level of 'direct' encouragement to the concern of international law. This temporal dimension, however, raises the critical question whether incitement to genocide can constitute a continuing crime.

The Nahimana Appeals Chamber rejected this characterisation, arguing that it contradicts incitement's inchoate nature. Once disseminated, they argue, the speech is uttered, and the incitement is thus deemed 'effective.'[320] Commentators, however, cautiously favour treatment of incitement as continuing.[321] Timmermann, drawing on Judge Shahabuddeeen's Dissenting Opinion in the *Nahimana* case, argues that rejecting incitement as a continuing crime 'is too rigid and artificial and does not take account of the inherently fluid and cumulative nature of incitement.'[322] Incitement, Shahabuddeen maintained, 'operates by way of the exertion of influence', which by itself 'is a function of the process of time.'[323] In her comprehensive work on incitement and international law, Timmermann proposes resolving this 'by drawing a distinction between an act being *treated* complete at the time when it is uttered, for the purpose of criminal punishment or sanctions on the one hand, and as having a continuing effect on the other.'[324] Even after 'completion', speakers remain responsible for subsequent effects.[325] While theoretically convincing and in line with the contemporary reality of incitement, it does question how long speech effects persist, and can be attributed to the initial expression.[326] Can an accumulation of messages still constitute a clear call for action? Without diminishing the narrow interpretation of incitement to genocide, international law can reasonably accommodate some crime continuity without lowering overall thresholds. The framework is rightfully rigid and demanding but should not be frozen in time to the point of failing its preventative purpose.

319 Section 1.3.1 'Technological Development and Digital Infrastructure'.

320 Gordon (n 1) 300-301; Schabas (n 256) 326; Orentlicher (n 284) 45; Timmerman (n 309) 825; Mohamed Elewa Badar, 'The Road to Genocide: The Propaganda Machine of the Self-Declared Islamic State (IS)' (2016) 16 International Criminal Law Review 361-411, 373; In *Bikindi*, the TC commented on temporality, noting that the speech 'must have been spoken at or near the time of the contextual violence', in *Prosecutor v Nahimana and others* (Judgment) ICTR-99-52-T (3 December 2003) paras 722-724.

321 Timmerman (n 50) 207, 209; Hartfort Guidelines (n 273) 40; Susan Benesch, 'The Ghost of Causation in International Speech Crime Cases' in Predrag Dojćinović (ed), *Propaganda, War Crimes Trials an International Law: From Speaker's Corner to War Crimes* (Routledge London 2012).

322 Timmerman (n 50) 209; *Prosecutor v Nahimana and others* (Appeals Chamber Judgment) ICTR-99-52-A (28 November 2007) Partly Dissenting Opinion of Judge Shahabuddeen, para 25.

323 *Ibid.*

324 *Ibid.* 210.

325 *Ibid.*

326 *Ibid.*

While the doctrinal requirement of *directness* continues to prioritise the clarity of the call to action in itself, contextual assessment standards have been developed for evaluating *direct incitement*, instead of the directness element in isolation. Among these are Benesch's 'reasonably possible consequences test',[327] and Wilson and Gillet's threshold that speech must 'significantly' increase the likelihood of genocide', proposed in the Hartford Guidelines. These standards focus on the likelihood that speech leads to genocide and enjoy widespread scholarly support.[328] Grounded in established practice and interpretation of the Genocide Convention and the *ad hoc* tribunals, Benesch's six-prong test is the most detailed and clarifies the existing 'direct and public incitement' standard by examining the content and context of the speech.[329] Despite criticism of rigidity, prosecution bias and definitional vagueness,[330] Benesch's framework forms the point of departure for positioning disinformation within existing parameters, examining:

1. Speakers and their influence
2. Audience and their susceptibility
3. Content and language, including falsehoods
4. Information environment
5. Dissemination method and technological amplifiers

Speaker

Similar to assessing incitement in hate speech, identifying incitement to commit genocide departs from the speaker and their authority relative to the audience.[331] While not categorically limited, Benesch notes that 'only some speakers *can* commit incitement to genocide – as much as others might wholeheartedly wish to do so – because only some have adequate influence or authority over the audience'[332] This connection between (perceived) authority and speaker influences is broadly evidenced beyond the legal field.[333] Within traditional incitement frameworks, the speaker's position

327 Benesch (n 64) 494-495; Benesch (n 321) 262-264.

328 Wilson (n 54) 224; Benesch (n 64) 494; Hefti and Ausserlandscheider (n 174) 1; Werle and Jeßberger (n 269); Van der Wilt (n 278) 48; La Mort (n 287) 4; Badar (n 320) 378-379, 380.

329 Benesch (n 64) 520-525; *Prosecutor v Nahimana and others* (Judgment and Sentence) ICTR-99-52-T (3 December 2003) paras 1000-1006.

330 Gordon (n 1) 275; Richard Ashby Wilson, 'Inciting Genocide With Words' (2015) 26 Michigan Journal of International Law (2015) 277-310; Badar (n 320) 379.

331 Mohamed Elewa Badar and Polona Florijančić, 'The *Prosecutor v Vojislav Sesilj:* A Symptom of the Fragmented International Criminalisation of Hate and Fear Propaganda' (2020) International Criminal Law Review 405-491, 480.

332 Benesch (n 64) 521; Hartfort Guidelines (n 273) 120.

333 Badar and Florijančić (n 331) 479 referencing HK Kelman, 'Violence Without Moral Restraint: Reflections on the Dehuminization of Victims and Victimizers' (1073) 29 Journal of Social Issues 1, 25-61; Stanley Milgram (ed), *Obedience to Authority: An Experimental View* (Harper and Row 1971); Jerry M Burger, 'Replicating Milgram: Would People Still Obey Today' (2009) 64 The American Psychologist 1, 1-11.

also affects the directedness of incitement: a rigid interpretation requires that the 'original genocidal message' must be delivered directly to a public audience without intermediaries.[334] As explained below, this is not a sustainable position in the modern information and media landscape. The position of speakers and speaker authority in discriminatory disinformation campaigns raises two critical questions: 1) how is authority established and exploited when spreading discriminatory disinformation in modern information environments; and 2) what are the legal implications of discriminatory disinformation passing through intermediaries?

Evolution of Authority in Digital Contexts of Discriminatory Disinformation

While government officials, politicians and other persons of public stature traditionally occupy positions of authority within the context of incitement to genocide, online global communities have created unprecedented influence relations between disinformation speakers and audiences. These developments occur outside the traditional stratagems and formal mechanisms of attaining a position of authority, power and influence, emphasising that the position of the speaker is predominantly relative.[335] What is decisive is the speaker's relative authority over a particular audience and ability to influence their thinking and actions through false discriminatory content.[336] Despite being relative, this authority in a disinformation context can be evaluated through objective factors, including the political prominence or the speaker's social position within majority or minority groups.[337]

Wilson's comprehensive work identifies perceived credibility and charisma as key risk factors in assessing incitement.[338] Discourse analysis confirms that (perceived) authority, and the credibility thereof directs the behaviour of the audiences.[339] Concerning incitement to genocide, discriminatory disinformation exploits this authority, manipulating perception through false narratives and creating a false impression of following via bots or fake personas to amplify traction. Disinformation is used to create artificial status perceptions and fabricates impressions of consensus – a *vox populi* that 'everyone is joining our bandwagon'.[340] Enabled by digital technologies, authority manipulation ultimately becomes self-reinforcing.

The type of authority exploited, however, differs. From trusted leaders or experts, varying from God to national heroes. Badar demonstrates how IS's

334 De Hemptinne (n 274) 403.
335 Section 5.4.4.4 'Speaker, Audience and Technology'.
336 Benesch (n 321) 266; Benesch (n 64) 520-521.
337 Timmerman (n 50) 217.
338 Wilson (n 54) 263; Hartfort Guidelines (n 273) 120.
339 *Ibid.* 228.
340 Oberschall (n 56) 172.

discriminatory disinformation calling for the destruction of the Yazidis,[341] claimed divine authority by pretending to speak in *Vox Dei*. Through the deliberate mischaracterisation of events and adversaries, the illusion is created that 'to deny the order of ISIS would be to deny the orders of Allah.'[342] Such claims of Prophetic representation are used to create 'an aura of legitimacy,' exploiting religious sentiment of obedience to god to advance discriminatory falsehoods.[343]

Disinformation is unlikely to amount to direct incitement without the involvement of a plurality of actors, challenging traditional doctrinal interpretations relying on a single (or limited number of) identifiable speaker(s) with clear authority and influence. Combined with online anonymity and questions about incitement as a continuing crime, this plurality hinders accountability. The alleged genocide against the Rohingya population in Myanmar exemplifies this challenge through the weaponisation of social media – primarily Facebook– by many different actors.[344] Dehumanisation and vilification campaigns against the Rohingya were coordinated and disseminated during and in the years prior to the large-scale violence, involved or were approved by government officials, nationalist political parties and politicians, prominent monks and academics. By leveraging authority, these actors manipulated public perception, inciting hostility towards the Rohingya population resulting in serious allegations of genocide,[345] while individual attribution remains elusive.

Intermediary "Speakers"

The Rohingya case demonstrates that rigidly interpreting *direct* incitement as requiring direct delivery of the 'original genocidal message', paralyses the legal framework.[346] Online manifestations of incitement to commit genocide, whether consisting of disinformation or not, raise fundamental questions about intermediary liability of media platforms,[347] and individual responsibility for subsequent dissemination of incitement leading to genocide. Accommodating this new reality, Timmerman and de Hemptinne suggest that the know-

341 Mohamed E Badar and Polona Florijančić, 'The Cognitive and Linguistic Implications of ISIS Propaganda: Proving the Crime of Direct and Public Incitement to Commit Genocide' in Predrag Dojčinović (ed), *Propaganda and International Criminal Law: From Cognition to Criminality* (Routledge 2020) 33-34, 39

342 *Ibid.* 34.

343 *Ibid.* 33.

344 UNHRC, 'Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar' (2018) UN Doc A/HRC/39/CRP.2, paras 1270, 1234, 1339-1352.

345 *Ibid.* 1324.

346 De Hemptinne (n 274) 403.

347 *Jane Doe v Meta Platforms*, Inc (f/k/a Facebook, Inc) (Class Action Complaint) Case No 21-CIV-06465 (Superior Court of the State of California for the County of San Mateo); Dan Milmo, 'Rohingya Sue Facebook for £150bn over Myanmar Genocide' The Guardian (6 December 2021).

ledge and the consent of the original speaker in subsequent dissemination should be determinative. In other words, did the speaker(s) know and agree to the 'revival of republication', and was the requisite intent still present? Did they still intend for the crimes to be committed at this point in time?[348]

'From a subjective standpoint', de Hemptinne argues, 'one can argue that if the individual was 'aware of the risk that his/her 'genocidal' message would be publicly disseminated through other persons and nonetheless willingly took that risk, he/she could be held responsible for incitement to commit genocide.'[349] An initially viable solution that constructively contributes to the debate on incitement as a continuing crime. It also addresses a critical issue in the prosecution of incitement to genocide in the digital era: exploitation of temporal remoteness as a defence strategy. Recognising the continuing nature of incitement – particularly in the context of public manipulation – prevents perpetrators from leveraging the time lapse between the creation of disinformation and genocidal acts to evade culpability.

Audience

The likelihood that the inciting speech translates into genocidal acts ultimately depends on audience reception and reaction.[350] Speakers' influence only constitutes direct incitement if the audience he or she reaches, first, understands the speech as a call for violence, and second, has the capacity to answer the call through action.[351] This prompts two preliminary questions in evaluating discriminatory disinformation: 1) who constitutes the audience, and 2) what determines their susceptibility?

Identifying the audience can trace the intent of the speaker and/or the foreseeable reach of the message: who is the initially intended audience from the perspective of the speaker(s) and/or which audience is most likely to react and respond to the speech?[352] Relevant indicators include language that addresses particular groups, existing ties between a speaker and a particular group, and whether – and by whom – previous violent calls by the speaker have been answered. A critical indicator of indirect incitement through discriminatory disinformation is whether the intended audience unequivocally under-

348 Timmerman (n 50) 210.

349 De Hemptinne (n 274) 403.

350 In the Streicher case, the IMT emphasised the importance of the connection between the speaker and the audience by relying on how the defendant's acts had affected the minds of other Germans as a form of mental causation, rather than focusing on how his words lead to concrete acts (physical causation) in Wilson (n 54) 31.

351 Benesch (n 64) 520-525; Francis M Deng, 'Contextualising the Prevention of Genocide' in André Nollkaemper and Julia Hoffman (eds), *Responsibility to Protect: From Principle to Practice* (Cambridge University Press 2021) 341.

351 Benesch (n 64) 521; Hartfort Guidelines (n 273) 121.

352 Gordon (n 1) 187; de Hemptinne (n 274) 402 ('inciters do not have to know who will be incited by their messages'); Benesch (n 321) 263.

stands a message.[353] ICTR and the ICTY proceedings have demonstrated that this estimation requires thorough knowledge of the language, culture, history and political context. Predicting this understanding is particularly challenging in online environments where audience anonymity and untransparent group structures often obscure both intended and actual audience identity. When the audience cannot be identified, estimating their potential reaction to discriminatory disinformation becomes virtually impossible.

An identifiable audience's likelihood of responding to discriminatory disinformation calling for violence depends on personal and in-group factors.[354] During ongoing conflicts, previous patterns of violence provide valuable indicators for predicting behavioural responses. Similar to findings on incitement to terrorism, specific socio-historical contextual factors also play an enabling role in the speech-beliefs-action triad.[355] The atrocities of the 20th century and the establishment of the Caliphate in Syria and Iraq broadly demonstrate that audience action in response to discriminatory disinformation calling for mass violence, including genocide, becomes more likely when preceded by other violent acts against the targeted group.[356] Established patterns of violence facilitate justifying rhetoric of further escalation through false discriminatory narratives.[357]

Predicting behavioural responses to discriminatory disinformation before violence manifests – the desired intervention points from a preventive perspective – presents greater challenges. Cognitive and behavioural experts have observed that 'under particular circumstances most people have the capacity for extreme violence and the destruction of human life.'[358] People 'can learn to commit […] atrocities against other people'[359] and may even 'by normal social structures' be brought to 'committing murderous ethnic cleansing.'[360] It requires, Oberschall concludes, 'conducive social conditions rather than monstrous people to produce heinous deeds'.[361] As illustrated in section 5.2.1.4, exposure to discriminatory disinformation influences individual suscept-

353 Predrag Dojćinović, 'In The Mind of The Crime: Mens Rea' in Predrag Dojćinović (ed), *Propaganda and International Criminal Law: From Cognition to Criminality* (Routledge 2020) 182.

354 Wilson (n 54) 229; Section 5.2.2 'Cognitive and Behavioural Mechanisms of Discriminatory Disinformation'; 5.4.4 'Incitement'.

355 Badar (n 320) 371; *Prosecutor v Bikindi* (Judgment) ICTR-01-72-T (2 December 2008) paras 247-250.

356 Susan Benesch (n 64) 522; International Panel of Eminent Personalities, 'Rwanda: The Preventable Genocide' (African Union, July 2000) Chapter 7.15.

357 Benesch (n 64) 522; Hartfort Guidelines (n 273) 120-121.

358 Staub (n 51) 62, 95.

359 Helen Fein, *Accounting for Genocide: National Responses and Jewish Victimisation during the Holocaust* (University of Chicago Press 1984) 33.

360 Michael Mann, *The Dark Side of Democracy: Explaining Ethnic Cleansing* (Cambridge University Press 2004) 9.

361 Oberschall (n 56) 183; Albert Bandura, *The Role of Selective Engagement in Terrorism and Counterterrorism* (Stanford University Department of Psychology 2004) 5, 24.

ibility to hateful rhetoric and may trigger behavioural responses.[362] Yet broader individual, political, economic and/or societal instability is necessary for disinformation to incite genocidal escalation. 'As a rule of thumb', Werle and Jeßberger explain, 'the more secure the internal peace among various groups in a society, the less defamatory and discriminatory statements should be interpreted as direct incitement to commit genocide.'[363]

Finally, the audience's estimated capacity to commit genocide is an underexposed contextual aspect of online incitement, including disinformation. This capacity may imply that the speech must reach a substantial audience,[364] or that there is a certain physical proximity between the audience and the targeted group. This digital information environment's speed, reach and scale make the potential audience virtually unlimited. While the individual capacity of this audience to act may be limited, their combined potential can surpass traditional thresholds of likelihood for genocidal acts. The risk, however, diminishes when the audience the speaker aims to incite is located on the other side of the world, geographically distant from potential victims.

Content, Language and Falsehoods

Whether speech is understood as a call to commit genocide depends on what has been said and why.[365] Incitement can take place even without explicit calls for violence,[366] particularly when rooted in falsehoods and manipulative linguistic patterns. Attempts to define direct incitement in terms of language have focussed on negative framing, referencing the ICTR's jurisprudence that 'a vague or indirect suggestion' is in any case not sufficient.[367] Instead, incitement must be 'specific enough to constitute instructions'[368] that 'specially provoke another to engage in a criminal act.'[369] The language used, however, may be 'nonetheless implicit.'[370] This position was reiterated by the International Law Commission, arguing that '[t]he element of direct incitement

362 *Ibid.* 192.

363 Werle and Jeßberger (n 268) 324; Benesch (n 64) 499.

364 Neema Hakim, 'How Social Media Companies Could be Complicit in Incitement to Genocide' (2020) 21 Chicago Journal of International Law 1, 96.

365 Gordon (n 1) 301.

366 Wibke K Timmermann, 'International Speech Crimes Following the Šešelj Appeal Judgment' in Predrag Dojinovi (ed), *Propaganda and International Criminal Law: From Cognition to Criminality* (Taylor & Francis 2019) 115, 118.

367 *Nahimana et al., v Prosecutor (*Appeals Chamber Judgment) ICTR-99-52-A (28 November 2007) para 692; *Prosecutor v Kajelijeli (*Judgment and Sentence) ICTR-98-44A-T (1 December 2003) para 85; *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T (2 September 1998) para 557; Albin Eser, 'Individual Criminal Responsibility' in Antonio Cassese, Paula Gaeta and J Jones (eds), *The Rome Statute of the International Criminal Court: A Commentary* (Oxford University Press 2002) Volume 1, 805.

368 *Prosecutor v Brdanin* (Trial Chamber Judgment) IT-99-36-T (1 September 2004) paras 468, 527, 662, 672.

369 *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T (2 September 1998) para 577.

370 *Ibid.*; *Prosecutor v Niyitegeka* (Judgment) ICTR-96-14-T (16 May 2003) para 431.

requires specifically urging another individual to take immediate criminal action rather than merely making a vague or indirect suggestion'.[371]

As Gordon rightfully points out, these efforts have, however, failed to provide a clear lexicon, a 'readily graspable phraseology', to coherently and consistently analyse the content of speech.[372] A paradox emerges: the most harmful and "successful" incitement techniques often lack directness in their literal meaning. Recognising these complexities, in *Prosecutor v. Akayesu*, the Trial Chamber determined that the content's cultural and linguistic meaning should be assessed from the perspective of the intended audience. A particular speech 'may be perceived as "direct" in one country, and not so in another, depending on the audience'.[373] Thus implicit incitement – often characteristic of discriminatory disinformation – may still be legally direct.[374] Textual meaning is not pre-determined by the speaker, but is constructed by the audience.[375] Linguistically, incitement thus can contain 'euphemistic, metaphorical or otherwise coded language'.[376]

Comparative research of incitement narratives preceding genocide reveals 'strikingly similar' patterns,[377] that are often visible in discriminatory disinformation campaigns. While genocidal violence is neither necessarily nor exclusively driven by hate,[378] common linguistic patterns include dehumanisation of victims that extensively employs false characterisations.[379] Accusations in the mirror – false claims that 'the victims-to-be are planning to commit atrocities against the genocidaires-to-be' and there are no other options to avert this –[380] and manipulation of historical accuracy are also prevalent.[381] Moral justifications of ongoing and future eliminationist action

371 International Law Commission, 'Report of the International Law Commission on the Work of its Forty-Eighth Session' (6 May–26 July 1996) UN Doc A/51/10, 26.

372 Gordon (n 1) 188, 284.

373 *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T (2 September 1998) para 556.

374 Schabas (n 256) 332; *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T (2 September 1998) para 577; *Prosecutor v Niyitegeka* (Judgment) ICTR-96-14-T (16 May 2003) para 431.

375 Frans Viljoen, 'Inciting Violence and Propagating Hate Through the Media: Rwanda and The Limits of International Criminal Law' (2020) 26 Obiter 1, 34-35.

376 De Hemptinne (n 274) 374; Florian. Jeßberger, 'Incitement (to Commit Genocide)' in Antonio Cassese (ed), *The Oxford Companion to International Criminal Justice* (Oxford University Press 2009) 373.

377 Benesch (n 64) 503-506.

378 Past atrocities have been carried out because of inescapable compliance to authority, out of peer solidarity or military loyalty, in Oberschall (n 56) 184-184.

379 Benesch (n 64) 503; Hartfort Guidelines (n 273) 120; Deng (n 351) 341; Nick Haslam, 'The Many Roles of Dehumanization in Genocide' in in Leonard S Newman (ed), *Confronting Humanity at its Worst: Social Psychological Perspectives in Genocide* (Oxford University Press 2019) 121.

380 Benesch (n 64) 504; Jeffrey Herf, *The Jewish Enemy: Nazi Propaganda During World War II and The Holocaust* (Harvard University Press 2006) 114.

381 Oberschall (n 56) 177; Hartfort Guidelines (n 273) 121; *Prosecutor v Bikindi* (Judgment) ICTR-01-72-T (2 December 2008) paras 254-255.

follows when the audience shows receptiveness to such hateful rhetoric.[382] Combined, 'the most effective condition' is the creation of 'threat messages that raise anxiety and fear' to induce 'a public demand for relief and action to reduce the threat.'[383] Social science research has theorised that extensive defamation, the use of derogatory terms and provocation of resentment against a specific group is indeed 'action-engendering'[384] and has a 'heightened impact in the context of genocide.'[385] However, to date, no coherent empirical evaluation exists relating to specific types of language and their inciting effects.[386]

Not all incitement to genocide – from a content perspective – involves disinformation. A simplified call to "exterminate them all", while morally reprehensible and legally inciting, contains no false or misleading information *per se* – although this nuance is of little practical relevance.[387] Conversely, more insidious rhetorical tools like accusations in the mirror ("they will destroy us") clearly qualify as disinformation content. Because the use of falsehoods is a recognised indicator of incitement, discriminatory disinformation itself constitutes a risk factor.[388] As Oberschall notes 'falsehoods and lies, from selective omission of facts, deliberate mischaracterisation of events and adversaries to out and out fabrication and lies' distinguish communications from propaganda.[389] Especially, Gordon adds, when falsity of a statement can be proven, this indicator may weigh significantly into the assessment.[390]

These linguistic complexities are further exacerbated by emerging technological challenges, in particular the prominence of AI-generated language.[391] Building on the findings from chapter one, synthetic media and AI-generated content is becoming increasingly sophisticated and AI systems can generate content that, for example, mimics regional dialects, cultural idioms and coded language. This can create – or at least contribute to – incitement that appears linguistically authentic to the target audience but maintains a degree of plaus-

382 Raul Hilburg, *The Destruction of the European Jews* (3rd edn, Yale University Press 2003) 1081.

383 Oberschall (n 56) 171 citing Carl Hovland *et al.*, *Communication and Persuasion* (Yale University Press 1963).

384 Tirell (n 6) 217.

385 Fein (n 359) 33; Werle and Jeßberger (n 268) 322 referencing Frank Chalk & Kurt Jonassohn, *The History and Sociology of Genocide* (Yale University Press 1990) 28; *Prosecutor v Nahimana and others* (Judgment and Sentence) ICTR-99-52-T (3 December 2003) para 1022.

386 Gordon (n 1) 188, 284; Wilson (n 54) 236.

387 Section 1.5.1.1 'Type of Information and Level of Veracity'.

388 Mendel (n 210) 60.

389 Oberschall (n 56)173.

390 Gordon (n 1) 298; Benesch (n 321) 260.

391 Fatemah Albader, 'Synthetic Media as a Risk Factor for Genocide' (2025) 16 Case Western Reserve Journal of Law, Technology & the Internet 200; Naman Anand, 'Of Code and Consequences: Assessing the Impact of Artificial Intelligence on International Criminal Law Norms Governing the Direct and Public Incitement to Genocide' (2004) 52 International Journal of Legal Information 2, 166-175.

ible deniability through semantic ambiguity, amplifying the already problematic and slightly paradoxical 'implicit yet direct' reality of incitement online.[392] AI-generated messages also complicate existing challenges of translation; the more sophisticated and advanced forms of 'implicit but direct' incitement become, the more difficult it will be for courts to adequately grasp their meaning. Directness can be lost or created in translation, which in turn affects the assessment of *mens rea* and likelihood of harm. These emerging technologies not only facilitate the weaponisation of false, inciting narratives, but also shield their detectability.

The standard from *Akayesu* thus needs to adapt to this dimension of language manipulation and the audience's perception of its origin and meaning. Practical assessment tools may include authentication protocols for distinguishing between human and AI-generate speech – Albadar, *inter alia*, discusses the possibilities of digital watermarks, metadata, and blockchain technology.[393] In addition, the evidentiary weight of translated as opposed to original-language incitement may need to be reconsidered, and – as argue elsewhere – courts will need to determine if the deployment of AI tools carries any probative value in terms of intent, as well as redefine their own role on how they will use technology to better detect these narratives.[394]

Information Environment

Beyond the speaker-audience-message triad, Benesch and others posit the functioning of the marketplace of ideas at a time when incitement took place, as a signalling factor.[395] The inciting effect is not merely the result of an accumulation of messages. The information ecosystem of these messages, including the availability of alternative views, pluralistic debate, and critical voices, determine their impact[396] The absence of these factors constitutes an early indicator of potential violent escalation. Is there is an uncensored and pluralistic exchange of ideas, or does a government or other authority exercise a media and information monopoly?[397]

The Yugoslav wars and the Rwandan conflict starkly illustrate the consequences of a dysfunctional information space, where disinformation and

392 Albander (n 391) 205-208, 218.

393 *Ibid*. 220.

394 Sections 1.4.3.2 'Standard of Best Available Science'; 1.5.1.3 'Intent (to Cause Harm)'; 2.4.2.2 'Tools, Features and Methods'.

395 Benesch (n 64) 523; Hartfort Guidelines (n 273) 120; Wilson (n 54) 263 (highlighting 'the speaker wields a monopoly on the means of communication or can censor and suppress information' as a risk factor).

396 Badar and Florijančić (n 331) 461 ('the totally of all state and non-state propaganda with essentially the same messaging and its constant repetition that surrounds the speech' that should be considered).

397 Jacqueline Fordyce, 'Genocide Denial, Disinformation, Armed Conflict: What Can Lawyers Do?' (JUSTICE Scotland Report, 20 June 2022) 6-7; Wilson (n 54) 230.

propaganda led certain groups to believe distorted versions of reality.[398] This reality was characterised by extensive and repeated accusations that the enemy was conducting, plotting, or desiring to commit the type of violence that the speakers intended to incite.[399] Observers of the trials following the genocide in Rwanda established this environment had 'direct causal effects on the perpetration of genocide.'[400] The head of the International Committee of the Red Cross mission in Croatia in the early 1990s similarly observed that 'the conflict [in Bosnia] was the first time I have seen such strong and effective propaganda on both sides. When you are talking to either side, they are absolutely convinced that they will be slaughtered by the other side.'[401] This 'paranoia propaganda', heavily infused with discriminatory disinformation,[402] triggered psychological effects nearly impossible to reverse.

In these circumstances, creating a 'coercive media environment'[403] favourable to genocide may itself constitute incitement.[404] The needed susceptibility to inciting speech develops gradually through consistent and repetitive exposure to manipulated information and genocidal rhetoric, without access to alternative sources. A psychological atmosphere emerges 'which allows genocide to flourish'.[405] Regime control and censorship over media, alongside spreading disinformation to discredit alternative sources, are important indicia of escalation. The contemporary online information environment's inability to provide the marketplace of ideas its corrective function,[406] introduces a new analytical dimension. Whether audiences have access to alternative views cannot be genuinely assessed without considering individual and group-based online information environments, as opposed to the traditional "marketplaces of ideas." Even when alternative information exists and is available, social media and its algorithmic structure can isolate individuals to the extent they believe no other views exist, creating a *de facto* 'message monopoly'.[407] This indirect manipulation particularly affects those already prone to seeking or

398 Badar and Florijančić (n 331) 462; Benesch (n 64) 523.
399 Kenneth L Marcus, 'Accusation in the Mirror' (2012) 43 Loyola University Chicago Law Journal 357-393, 359; *Prosecutor v Nyiramasuhuko and others* (Judgment and Sentence) ICTR-98-42-T (24 June 2011) para 6026.
400 Markus (n 409) 378; Catharine MacKinnon, 'International Decisions: Prosecutor v. Nahimana, Barayagwiza & Ngeze' (2004) 98 American Journal of International Law 325, 330.
401 Anthony Oberschall, 'Vojislav Seselj's Nationalist Propaganda: Contents, Techniques, Aims and Impacts 1990-1994' (UN International Criminal Tribunal for the Former Yugoslavia, 4 January 2005) exhibit no P00005, 10.
402 Badar and Florijančić (n 331) 467.
403 Carol Pauli, 'Killing the Microphone: When Broadcast Freedom Should Yield to Genocide Prevention' (2010) 61 Alabama Law Review 4, 679.
404 Timmerman (n 220) 269.
405 Oberschall (n 56) 173, 189-190; Badar (n 320) 362; Wilson (n 54) 285; Dodd (n 245) 1; Benesch (n 321) 263.
406 Section 1.4.1.1 'Weaponisation of Protected Speech'.
407 The term comes from Oberschall (n 56) 174.

being exposed to inciting expressions and disinformation, including individuals with extremist worldviews or conspiracy-related tendencies.

Dissemination and Technological Amplification

Connected to the information environment, the mode of transmission or channel of communication has emerged as an indicator of incitement.[408] Benesch suggests examining whether 'the speech [was] transmitted in a way that would reinforce its capacity to persuade, e.g. via a media outlet with a particular influence, or set to compelling music.'[409] While the ICTR, *inter alia*, underscores that no *a priori* limitations on the mode of transmission or channel of communication exist,[410] different means and methods increasingly influence reach, persuasiveness and pervasiveness. Although this differentiation does not have an exclusionary effect, it should not be left ignored due to its influence on the whole contextual assessment. The choice of channel, for example, directly affects the speakers-audience relationship, the speaker's perceived authority and the audience's susceptibility. During the Rwandan genocide, 'the radio was akin to the voice of God […],'[411] while in Myanmar's Internet-reliant media landscape, the inciting effect on social media was particularly impactful, because 'Facebook is the internet'.[412] Moreover, the medium may indicate the level of governmental involvement, perceived trustworthiness and the potential for early intervention.

Beyond the discussion on the impact of AI and deepfake technology explored in section 5.4.4.2, legal and scholarly discourse has yet to fully explore these nuances. As one of the few commentators, Pauli 'distinguishing the impact of a book from that of a bullhorn', examines several ECtHR cases that seem to indicate a 'more forgiving' attitude towards literary works compared to mass media, attributing this to differences in availability, circulation and

408 Gordon (n 1) 299-300; Pauli (n 403) 685-686.

409 Benesch (n 321) 264.

410 Incitement comprises 'directly provoking the perpetrator(s) to commit genocide, whether through speeches, shouting or threats uttered in public places or at public gatherings, or through the sale or dissemination, offer for sale or display of written material or printed matter in public places or at public gatherings, or through the public displays of placards or posters, or through any other means of audiovisual communication'; Schabas notes that these words 'have been cited with approval' by other TC's of the ICTR, its AC and the Supreme Court of Canada, in Schabas (n 256) 332 citing *Prosecutor v Muvunyi* (Judgment) ICTR-2000-55A-T (12 September 2006) para 502;*Prosecutor v Kajelijeli* (Judgment and Sentence) ICTR-98-44A-T (1 December 2003) para 853; *Prosecutor v Niyitegeka* (Judgment and Sentence) ICTR-96-14-T (16 May 2003) para 431; *Nahimana and others v Prosecutor* (Appeals Chamber Judgment) ICTR-99-52-A (28 November 2007) paras 698-702; *Mugesera v Canada* (Minister of Citizenship and Immigration) [2005] 2 SCR 100, paras 87, 9.

411 Benesch (n 64) 521 citing Romeo Dallaire, *Shake Hands With the Devil* (Arrow Books Ltd 2005) 272.

412 Emma Irving, 'The Role of Social Media is Significant': Facebook and the Fact-Finding Mission on Myanmar' OPINIO Juris (7 September 2018).

associated risk to national security and public order.[413] Heightened scrutiny should, *inter alia*, be applied to channels which have a history of engaging with false and inciting information, or belong to actors with a history of inciting violence. Using communication channels and platforms known to have distortive information dissemination structures equally weighs heavily into the assessment. If availability and circulation influence the likelihood of incitement, the technological infrastructure and algorithmic programming underpinning the dissemination of discriminatory disinformation warrants closer scrutiny. Online interfaces, designed to promote emotionally charged content facilitate rapid proliferation of inflammatory rhetoric and hate propaganda.[414] As Deng notes, this 'sets a tone of impunity, even if it does not amount to incitement to genocide itself'.[415]

5.5.2.3 Mens Rea

Genocide is ultimately defined by genocidal intent – the intent to destroy the targeted group in whole or in part. For incitement to genocide, a dual intent requirement applies: the speaker must have the intent to provoke another individual to commit genocide and have intent to commit genocide itself.[416] The applicable standard for 'intent to incite' is ambiguous, with courts and tribunals thus far having failed to uniformly clarify it. While some scholars argue that establishing intent is 'no practical difficulty,'[417] it remains a significant prosecution obstacle.[418] In relation to, *inter alia*, Article 25(3) of the Rome Statute, commentators emphasise the inciter's knowledge that 'he was acting publicly and that his acts had a direct inciting effect on others persons' alongside genocidal intent.[419] Incitement's *mens rea* thus captures 'a desire on the part of the perpetrator to create by his actions a particular state of mind necessary to commit such a crime on the minds of the person(s) he is so engaging.'[420] Additionally, the 'inciter must know that the audience will understand his/her call as one to commit genocide.'[421] While genocidal intent

413 Gordon (n 1) 300; Pauli (n 403) 686 citing *Zana v Turkey* App no 18954/91 (ECtHR, 25 November 1997).
414 Oberschall (n 56) 172-173; Benesch (n 64) 524; Hartfort Guidelines (n 273) 120.
415 Deng (n 351) 342.
416 *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T (2 September 1998) para 560.
417 Schabas (n 256) 326 ('the *mens rea* generally obvious enough from the content of the message').
418 Paul Behrens, 'A Moment of Kindness' in Ralph Henham and Paul Behrens (eds), *The Criminal Law of Genocide* (Ashgate 2007)127.
419 Timmerman (n 50) 219; Albin Eser, 'Individual Criminal Responsibility' in Antonio Cassese *et al.*, *The Rome Statute of the International Criminal Courts: a Commentary* (Oxford University Press 2002) Vol I, 767, 806; Kai Ambos, 'What Does 'Intent to Destroy' Mean?' (2005) 91 International Review of the Red Cross 876, 833-858.
420 *Prosecutor v. Akayesu* (Judgement) ICTR-96-4 (September 2 1998) para 560.
421 De Hemptinne (n 274) 405.

is a prerequisite, it need not be the sole motivation. Personal motives may coexist without precluding genocidal intent.[422]

Publicly inciting discriminatory disinformation spread with intent to incite genocide unequivocally constitutes prohibited speech. Even when it does not meet this threshold, disinformation still indicates malevolent intent in broader incitement campaigns. Either because the propagation of falsehoods is generally deemed indicative of intent ('language and falsehoods') or because the intent component inherent in disinformation implies knowledge of potential harmful consequences ('knowledge and circumstances).

Language and falsehoods

The language used is a prominent indicator of intent,[423] allowing intent to be inferred from racist and ethnic hate propaganda.[424] Propaganda, Jacques Ellul observed, 'is necessarily a declaration of one's intentions.'[425] Instrumentalising disinformation 'can reveal or at least indicate the intent hidden behind the entire propaganda mechanism.'[426] Both the ICTY and the ICTR jurisprudence have subscribed to this view. The Akayesu Trial Chamber noted that 'the propaganda campaigns conducted before and during the tragedy by the audiovisual media, for example "Radio Television des Milles [*sic*] Collines" (RTLM), or the print media, like the Kangura newspaper' formed an important factor in finding that 'it was indeed the Tutsi who were targeted.'[427] Oberschall's discourse analysis on the Šešelj trial further evidences how propaganda and the use of falsehoods 'were calculated to deceive and manipulate,' and – though not all completely fabricated – it clearly illustrates Šešelj's advocacy for eliminationist actions.[428] Similarly, Badar identifies ISIS publications and broadcasts as textbook cases of propaganda

422 *Ibid.* 399; *Prosecutor v Jelisić* (Appeals Chamber Judgment) IT-95-10-A (5 July 2001) para 49; *Prosecutor v Krnojelac* (Appeals Chamber Judgment) IT-97-25-A (17 September 2003) para 102; *Prosecutor v Ntakirutimana* (Appeals Chamber Judgment) ICTR-96-10-A and ICTR-96-17-A (13 December 2004) paras 302-304; *Prosecutor v Niyitegeka* (Appeals Chamber Judgment) ICTR-96-14-A (9 July 2004) paras 48-54; *Prosecutor v Simba* (Appeals Chamber Judgment) ICTR-01-76-A (27 November 2007) para 269.

423 *Prosecutor v Nahimana and others* (Judgment and Sentence) ICTR-99-52-T (3 December 2003) para 1001.

424 Timmerman (n 50) 220.

425 Badar (n 320) 368; Jacques Ellul, *Propaganda: The Formation of Men's Attitudes* (Vintage 1973) 56.

426 Predrag Dojćinović, 'Word Scene Investigations: Towards a Cognitive Linguistic Approach to the Criminal Analysis of Open Source Evidence in War Crimes Cases' in Predrag Dojćinović (ed), *Propaganda, War Crimes Trials an International Law: From Speaker's Corner to War Crimes* (Routledge London 2012) 72.

427 *Prosecutor v Akayesu* (Judgment) ICTR-96-4-T (2 September 1998) para 123.

428 Oberschall (n 56) 178; Edina Bećirević, 'The Issue of Genocidal Intent and Denial of Genocide A Case Study of Bosnia and Herzegovina' (2010) 24 East European Politics and Society 4, 480-502.

containing false and misleading information that 'clearly demonstrate the requisite *mens rea* for direct and public incitement.'[429]

Disinformation is characterised by the intent to instrumentalise the false or misleading nature of the information to cause harm, making its dissemination *per definitem* a declaration of someone's intentions.[430] Recalling the differentiation between mis- and disinformation,[431] disinformation exceeds the "mere" use of false facts in the dissemination of propaganda, enhancing its indicative value. However, not all disinformation, discriminatory or otherwise, carries equal probative value. False information directly linked to violent rhetoric or eliminationist solutions more strongly reflects an intent to incite genocide than abstract falsities focussed on group characteristics. In this assessment, both the form and scale of disinformation matter. While a single piece of disinformation is unlikely to evidence genocidal intent, disinformation campaigns and operationsmay. Such large-scale employment of falsehoods implies coordination, strategy and allocations of resources, and should receive greater weight in determining intent to incite genocide.

Knowledge and circumstances

Beyond the falsity of the information and linguistic features, the intent to incite may be inferred from the presence of disinformation as such.[432] While international tribunals and commentators have approached contextual scrutiny differently,[433] the presence of propaganda has been recognised as circumstantial element indicating the existence of genocidal patterns and/or the intent to destroy since the early days of the Genocide Convention's negotiations.[434]

Discriminatory disinformation that falls short of incitement but aligns with hate speech implies the speaker's knowledge of potentially harmful consequences. Importantly, knowledge does not necessarily imply intent and therefore the intent-requirement cannot be fulfilled with reference to knowledge of the perpetrators alone. Conversely, intent implies knowledge, and disinformation always implies intent. Even if not directed towards inciting genocide – or it cannot be proven that it is – the intent to, *inter alia,* incite discrimination, hostility or violence, may constitute circumstantial evidence. Not as a decisive,

429 Badar and Florijančić (n 331) 50.

430 Hefti and Ausserlandscheider (n 174) 32; Mendel (n 210) 60 referencing *Walendy v. Germany* App no 21128/93 (ECtHR, 23 September 1992); Christopher Scott Maravilla, 'Hate Speech as a War Crime: Public and Direct Incitement to Genocide in International Law' (2008) 17 Tulsa Journal of International and Comparative Law 113, 141.

431 Section 1.5.1.1 'Type of Information and Level of Veracity'.

432 De Hemptinne (n 274) 405; *Mugesera v Canada (Minister of Citizenship and Immigration)* [2005] 2 SCR 100, 2005 SCC 40.

433 For an overview see Gordon (n 1) 10 and Timmerman (n 50) 220.

434 Nasour Koursami, *Contextual Elements of the Crime of Genocide* (Dissertation, University of Edinburgh 2016) 47, 81 citing Mr Morozov (Union of Soviet Socialist Republics) UN Doc E/AC.25/SR.26 (1948); Predrag Dojčinović, 'The Forensification of Propaganda in Epic Poetry' in Fiana Gantheret *et al.* (eds), *Art and Human Rights* (Edward Elgard 2023) 143.

but as a contributing factor. Disinformation's probative value is further enhanced when forms of disinformation containing similar messages and (false) information have previously led to outbreaks of violence. In addition, if the actors disseminating information have a history of engaging with inciting violence, this presumes knowledge of potential consequence. These patterns and experiences, Benesch argues, 'put speaker and audience on notice that such speech can indeed lead to violence, providing evidence of the speaker's intent and increasing the dangerousness of the speech.'[435] Actors with a history of creating and disseminating harmful and untruthful information, should be thus subject to additional scrutiny.[436]

Scale and coordination of disinformation campaigns and operation further indicate *mens rea.* Regardless of whether incitement to genocide can be of a continuous nature, the continuous nature of disinformation is indicative of the required *mens rea*. The exposure to anti-Rohingya propaganda campaign in Myanmar exemplifies this: Myanmar military personal flooded Facebook with disinformation stories and campaigns for over half a decade, intentionally and systematically inducing fear and resentment, and subsequently gaining support for 'a textbook example of ethnic cleansing'.[437] Researchers demonstrated how hundreds of military personnel created troll accounts, spreading false and inflammatory news, shutting down criticism and creating violent disinformation narratives faking evidence of Rohinya-perpetrated massacres, imminent 'jihad attacks' and personal denigration of democratically elected leaders.[438] Due to temporal distance between these action and the alleged mass violence, and the unlikelihood that the majority of these stories in isolation would reach the threshold of direct incitement, the probative value of these patterns of discriminatory disinformation lies in evidencing an intent to incite genocide.

5.5.2.4 Interim Conclusion

In sum, disinformation, when public, directly inciting and accompanied by genocidal intent, is unequivocally prohibited under international criminal law. However, traditional doctrine on incitement to genocide does encounter

435 Benesch (n 321) 264.

436 Szakacs and Bognar (n 21) vii; Deng (n 351) 343; Anne Merlan, 'How Covid Conspiracy Theories Led to an Alarming Resource in AIDS Denialism' (7 August 2024) MIT Technology Review; Matthew Parnell and Mary Stuckey, 'Holocaust Distortion During the Global Pandemic: An Exercise in Anti-Democratic Demagoguery' (2023) 30 Journal of the European Institute for Communication and Culture 4, 1-17.

437 United Nations, 'UN Human Rights Chief Points to 'Textbook Example of Ethnic Cleansing' in Myanmar' (United Nations Press Release, 11 September 2017).

438 Paul Mozur, 'A Genocide Incited on Facebook, With Posts From Myanmar's Military' The New York Times (15 October 2018); Maun Zarni and Alice Cowley, 'The Slow-Burning Genocide of Myanmar's Rohinya' (2014) 23 Washington International Law Journal 3, 684-754.

significant challenges when applied to online communications, new spheres of influence and global audiences. The protracted nature of discriminatory disinformation conflicts with traditional juridical conceptions of incitement as a distinct, temporally bounded, act. Contemporary incitement through false and misleading information manifests as cumulative campaigns across platforms involving multiple actors. Successful incitement is a continuing process – a reality the doctrine should centralise to uphold it preventive rationale. Digital communications have also complicated the assessment of 'public incitement', as platform structures obfuscate traditional boundaries between public and private spheres. Similarly, the directness faces challenges from coded language, algorithmic amplification, and implicit messaging that are not only complex to decipher for outsiders, but increasingly complex to identify in the first place.

Even when disinformation does not rise to the level of incitement to genocide, it influences the contextual factors used to determine whether speech constitutes incitement. Creating coercive media environments, instrumentalising algorithmic amplification, and drowning out alternative narratives all foster conditions conducive to genocidal violence. Disinformation campaigns shape the audience's susceptibility to inciting messages, manipulate perceptions of speaker authority, and distort the information landscape in ways that traditional incitement frameworks do not fully capture. The Myanmar case in particular demonstrates how systematic discriminatory disinformation facilitates psychological environments where genocide becomes possible. Perhaps most significantly, discriminatory disinformation serves as a powerful indicator of genocidal intent. The intentional instrumentalisation of falsehoods, particularly in coordinated campaigns, implies planning and strategy relevant to proving *mens rea*. Pattern and content analysis of disinformation campaigns containing accusations in the mirror or fabricated threats, provides critical circumstantial evidence of intent to incite violence against protected groups. Recognising disinformation as both potential incitement and key evidence of intent advances legal approaches to preventing and punishing genocide in the modern information environment.

5.5.3 Genocide Denial

Having demonstrated how disinformation can catalyse genocide through incitement, this section examines disinformation's role in the aftermath of atrocities through denial, completing a cycle that both enables and obscures mass violence. Regulating genocide denial, however, represents one of the most contested domains in legal scholarship. Labelled 'the most controversial

issue related to freedom of expression' in Europe,[439] 'denialism' manifests throughout various phases of atrocities – before, during and after their commission.[440] Corresponding with these time frames, disinformation narratives shift: systematic starvation becomes portrayed as 'disease' or 'famine', and structural denial of basic needs as a 'failure of the international community to provide needed relief'.[441] Simultaneously, post-atrocity strategies include destroying historical records or dismissing them as propaganda, discrediting reporting or investigative mechanism, and exploiting "definitionalism" – sowing doubts to contest whether certain action constituted genocide.[442]

Stanton's influential 'eight stages of genocide' from 1996 concludes that 'every genocide is followed by denial.'[443] The digital era has witnessed an exponential increase in denialist content online, particularly trivialising propaganda and disinformation.[444] This increase has reinvigorated debates on balancing freedom of expression and protecting victims, preserving historical truth, and fostering societal reconciliation.[445] The contentious nature of this issue was starkly illustrated by the 2024 UNGA Resolution condemning the denial of the Srebrenica genocide as a historical event; 84 States voted in favour, 19 against and 68 abstained.[446]

Denialist disinformation as a form of discriminatory disinformation encompasses the distortion or manipulation of historical truths and factual events

439 Ludovic Hennebel and Thomas Hochmann (eds), *Genocide Denials and the Law* (Oxford University Press 2011) x1iii.

440 Stanton (n 251); Colin W Leach *et al.*, 'Moral Immemorial: The Rarity of Self-Criticism for Previous Generations' Genocide or Mass Violence' (2013) 69 Journal of Social Issues 1, 34-53.

441 Clotilde Pégorier, 'Speech and Harm: Genocide Denial, Hate Speech and Freedom of Expression' (2018) 18 International Criminal Law Review 97-126, 104-105; Gregory Stanton, 'The Ten Stages of Genocide' (Genocide Watch 2017); Edina Becirevic, 'The Issue of Genocidal Intent and Denial of Genocide' (2010) 24 East European Politics and Societies, 487-492.

442 Israel W Charny, 'Templates for Gross Denial of a Known Genocide: A Manual' in Israel Charny (ed), *The Encyclopaedia of Genocide* (ABC Clio 1999) 168; Clotilde Pégorier, *Ethnic Cleaning: A Legal Qualification* (Routledge 2013) 17-33.

443 Stanton (n 251).

444 Michael Whine 'Expanding Holocaust Denial and Legislation Against it' in Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (Oxford University Press 2009) 539; Piotr Bakowski, *Holocaust Denial in Criminal Law: Legal Frameworks in Selected EU Member States* (Brief European Parliamentary Research Service, January 2022) 2.

445 Hennebel and Hochman (n 439) xiii; Rezarta Bilali, Yeshim Iqbal and Samuel Freel, 'Understanding and Counteracting Genocide Denial' in Leonard S Newman (ed), *Confronting Humanity at its Worst: Social Psychological Perspectives in Genocide* (Oxford University Press 2019) 284; Paul Behrens, 'Genocide Denial and the Law: A Critical Appraisal' (2015) 21 Buffalo Human Rights Law Review 2, 32; Rob Kahn, 'Can the Law Understand the Harm of Genocide Denial' in Roland Moerland *et al.*, *Denialism and Human Rights* (Intersentia 2018) 230.

446 UNGA, '78th Session, 82nd Plenary Meeting' (23 May 2024) UN Doc GA/12601; UNGA 'International Day of Reflection and Commemoration of the 1995 Genocide in Srebrenica' (20 May 2024) UN Doc A/78/L.67/Rev.1.

to diminish the responsibility of perpetrators,[447] and/or to incite discrimination or violence.[448] This conceptualisation bridges established scholarship and regulation on denialism with disinformation, focussing on two key elements: harmful intent and factual distortion. Scholarship on denialism reflects this by commonly distinguishing between 'aggravated' (explicitly targeting particular groups, e.g. claims that '"the Jews" invented the hoax of the Holocaust to exploit Germany') and 'bare' denial (factual contestation without explicit targeting, e.g. 'no gas chambers were used during WWII').[449] This distinction matters legally, as aggravated denial aligns better with the structure of unlawful hate speech regulation.[450] Similarly useful is Cohen's typology of 1) literal denial as full factual negation ('the Holocaust did not happen'); 2) interpretative denial by accepting facts but reframing meaning; and 3) implicatory denial which trivialises implications or consequences.[451]

It encompasses all three forms when characterised by an intent to cause harm through denying the existing or characteristic features of established atrocities.[452] This intent-focused approach averts critiques that denial laws risk becoming 'dangerous political intrusion into the domain of objective historical inquiry'. By focussing on intent, courts distinguish between permissible and impermissible forms of genocide denial, and separate mis- from disinformation.[453] Without this distinction, international scholars and lawyers could face unjust accusations of denial or revisionism when examining genocide's legal parameters, in theory or practice. Denialist disinformation shows a great resemblance to other types of disinformation. It often occurs within broader conspiracy theories and racist ideology. When, *inter alia*, appropriated by terrorist organisations, it creates a conceptual overlap with terroristic disinformation (chapter four),[454] Similarly, targeted forms may coincide with the

447 Bilali, Iqbal and Freel (n 445) 286; Roger Schmith: 'Legislating Against Genocide Denial: Criminalizing Denial or Preventing Free Speech?' (2010) 4 Journal of Law and Public Policy 2, 128; Marko Milanovic, 'State Lies as Violations of Human Rights' (5 May 2025) Human Rights Quarterly Forthcoming, 60.

448 Shannon Fyfe, 'Tracking Hate Speech Acts as Incitement to Genocide in International Criminal Law' (2017) 30 Leiden Journal of International Law 525-548, 545.

449 Ludovic Hochmann, 'The Denier's Intent' in Ludovic Hennebel and Thomas Hochmann (eds), *Genocide Denials and the Law* (Oxford University Press 2011) 280; Kahn (n 445) 223.

450 Kahn (n 445) 222-223; Hochmann (n 449) 280.

451 Stanley Cohen, *The States of Denial: Knowing About Atrocities and Suffering* (Polity 2001) 7-9.

452 Hennebel and Hochman (n 439) xix citing Carole Vivant, *'L'Historien Saisi Par le Droit'* (Contribution à l'étude des droit de l'histoire, Paris 2007) 417 ("purest form" of denial is 'an expression contesting the existing of the crime of a characteristic feature of the crime').

453 *Bona fide* mistakes qualify as misinformation, based on Hochmann (n 445) 285, 305; Kahn (n 445) 227; Paolo Lobba, 'Criminalizing Negationism Beyond the Holocaust' (2013); David Fraser, 'On the Internet, Nobody Knows You're a Nazi' Ivan Hare and James Weinstein (eds), *Extreme Speech and Democracy* (Oxford University Press 2009) 513.

454 Alida Skiple, 'Whitewashing White Power: a Rhetorical Political Analysis of the Parliamentary Ambition of the Nordic Resistance Movement in Sweden' (2023) Journal of Political Ideology 1-19.

scope of defamatory disinformation (chapter three) or when instrumentalised towards interference in the internal affairs of another State, it may constitute prohibited intervention (chapter two).

The question is not whether genocide denial causes harm, but whether restricting it is compatible with freedom of expression.[455] Denialism inflicts direct psychological harm on victims of genocide and their descendants.[456] It undermines victims' suffering, devaluates their experience and perpetuates intergenerational trauma.[457] As Stanton notes, '[d]enial is a continuation of a genocide, because it is a continuing attempt to destroy the victim group psychologically and culturally, to deny its members even the memory of the murders of their relatives.'[458] It also threatens historical truth, undermining societal reconciliation processes and eroding collective understanding of past atrocities essential for preventing future violence.[459] Empirical research likewise demonstrates that denial fuels 'moral disengagement' – the gradual process through which people are desensitised to violence against target groups.[460] Through this mechanism, denialism 'carries the seed for the commission of further international crimes',[461] creates enabling conditions for renewed violence.

International legal framework

'Genocide denial' lacks an explicit definition in international law and courts and human rights mechanisms have addressed it ambivalently. The ECtHR has held that denial of atrocities can fall outside the scope of freedom of expression,[462] repeatedly finding complaints of Holocaust deniers – challenging undue restriction of their freedom of expression – inadmissible.[463] Yet in *Perincek*, declaring the Armenian genocide was 'an international lie' was

455 Milanovic (n 447) 57.
456 Bilali, Iqbal and Freel (n 445) 284.
457 Hennebel and Hochman (n 439) 32.
458 Stanton (n 251).
459 Bilali, Iqbal and Freel (n 445) 284.
460 *Ibid.* 287; Albert Bandura, 'Moral Disengagement in the Perpetration of Inhumanities' (1999) 3 Personality and Social Psychology Review 3, 193-209; Albert Bandure, 'Selective Moral Disengagement in the Exercise of Moral Agency' (2002) 31 Journal of Moral Education 2, 101-119.
461 Behrens (n 445) 32; Fordyce (n 392) 6 ('[g]enocide denial and disinformation have been used to eliminate the capacity of the dominant group to feel empathy towards the targeted group, lessening the potential for internal dissent against genocidal policies'); Israel Charny, 'Innocent Denials of Known Genocides: A Further Contribution to a Psychology of Denial of Genocide' (2000) 1 Human rights Review 3, 15-39; Karen Etlis, 'A Constitutional Right to Deny and Promote Genocide? Preempting the Usurpation of Human Rights Discourse Towards Incitement From a Canadian Perspective (2008) 9 Cardozo Journal of Conflict Resolution 463-477.
462 Aswad and Kaye (n 142) 177-179; Lerner (n 133); Fyfe (n 445) 545.
463 Pégorier (n 441) ft. 9 for an overview.

held to violate the applicant's freedom of speech.[464] The Human Rights Committee General Comment No 34 notes that penalisation of 'opinions about historical facts' violated freedom of expression and opinion,[465] but in *Faurisson v. France*, the Committee held that the criminal prosecution for Holocaust denial did not amount to a violation.[466] The Committee echoed the ECtHR that negation or revision of 'clearly established historical facts – such as the Holocaust – is not covered by the protected speech.[467] This follows the Court's distinction from *Lingens v. Austria* 'between facts and value-judgement'.[468] The ICERD also recommends that 'public denials or attempts to justify crimes of genocide and crimes against humanity, as defined by international law, should be declared as offences punishable by law' if they 'clearly constitute incitement to racial violence or hatred.'[469] In *Jewish Community of Osla v. Norway*, the Committee recognised that statements of racial superiority can qualify as incitement depending on context, and Holocaust denial may reach this threshold.[470] To the contrary, in GR 35, the Committee reiterated the ICCPR's position that '[o]pinions about historical facts' should not be prohibited or punished.[471]

Criminal prohibition of genocide denialism exists primarily in Europe, with Colombia, Israel and Rwanda as notable exceptions.[472] Many other States, however, do address denial of genocide as 'a kind of hidden hate speech,' depending on the circumstances and how the speech is presented and disseminated.[473] National laws cover various expressions, including bare denial,[474] minimisation, justification and approval of genocide,[475] sometimes requiring discriminatory intent.[476] Like human rights mechanisms, national frameworks tend to distinguish between Holocaust denial and denial of other genocides.

---

464 *Perinçek v Switzerland* App no 27510/08 (ECtHR, 15 October 2015) paras 114-115.
465 General Comment No 34 (n 80) para 49.
466 *Faurisson v. France* (n 136).
467 *Lehideux and Isorni v France* App no 55/1997/839/1045 (ECtHR, 23 September 1998) para 47.
468 *Lingens v Austria* App no 9815/82 (ECtHR, 8 July 1986) para 46.
469 ICERD General Recommendation 35 (n 89) para 14.
470 *The Jewish Community of Oslo and Others v Norway* (n 199) para 10.4; Thornberry (n 92) 292; UNGA 'Report of the Committee on the Elimination of Racial Discrimination' (1997) UN Doc A/52/18 para 217 (ICERD sided with Germany's efforts of legally prohibiting genocide denial, though it considered Germany's legal framework 'too restricted' due to limiting applicability to the Holocaust).
471 ICERD General Recommendation 35 (n 189); General Comment No 34 (n 80) para 49.
472 William R Pruitt, 'Understanding Genocide Denial Legislation: A Comparative Analysis' (2017) 12 International Journal of Criminal Justice Sciences 2, 271.
473 Hennebel and Hochmann (n 435) x1v; Behrens (n 445) 32.
474 Including statements without any explicit accusations against a group, in Hochmann (n 445) 281.
475 Pruitt (n 472) 271.
476 *Ibid.* 273.

Within international legal frameworks, denialist disinformation that qualifies as aggravated denial clearly falls within the parameters of unlawful hate speech.[477] These expressions target particular groups based on identity with intent to further discrimination, hostility or violence, carrying likelihood of inflicting harm.[478] Even if denialist disinformation does not rise to the level of incitement, it may still a constitute permissible restriction to freedom of expression under Article19(3) ICCPR if necessary 'for respect of the rights or reputations or others' or to protect 'public order' and conforms to the tripartite test outlined in chapter one.

Beyond freedom of expression, criminalisation of genocide denial impacts fair trial guarantees, the rights of defendants and judicial proceedings. In 2010, ICTR defence counsel Peter Erlinder was arrested in Rwanda for alleged genocide denial while representing his client, undermining the presumption of innocence.[479] The ICTY addressed denialist disinformation in the context of procedural disruption. The 2002 report by the former Republika Srpska Government Bureau for Relations with the ICTY titled 'Report about the Case Srebrenica', denying the occurrence of the massacre and claiming that the International Committee of the Red Cross in documenting the atrocities had manipulated and fabricated evidence,[480] was condemned as a 'one of the worst examples of revisionism.'[481] The backlash by the international community was unequivocal, echoing the sentiment expressed by the High Representative for Bosnia in Herzegovina that it was 'tendentious, preposterous and inflammatory' and 'so far from the truth as to be almost not worth dignifying with a response.'[482]

Disseminating denialist disinformation also has implications for individual criminal responsibility. Before the ICC Pre-Trial Chamber in the *Mbarushimana*, the Prosecution claimed that the defendant had contributed to the FDLR's crimes by issuing 'several press releases on behalf of the organisation in the aftermath of the operations, systematically denying any responsibility of the group' as well as 'shrewdly portraying the FDLR as an actor seeking peace

477 Section 5.4 'Unlawful Hate Speech in International Law'; 'Hennebel and Hochmann (n 435) xix.

478 Fyfe (n 448) 545.

479 Behrens (n 445) 38, 39; The ICTR called for Erlinger's immediate release in Office of the Registrar of the ICTR, 'Note Verbale to the Ministry of Foreign Affairs and Cooperation of the Government of Rwanda' (15 June 2010) ICTR/TO/06/10/175; the American Bar Association called upon Rwanda to comply with the UN Basic Principle on the Role of Lawyers (UNGA 'Basic Principles on the Role of Lawyers' (7 September 1990) UN Doc A/CONF.144/28/Rev.1 Article16(a)) which imposed an obligation to ensure that lawyers are 'able to perform all of their professional functions without intimidation, hindrance, harassment or improper interference.')

480 Behrens (n 445) 39.

481 *Prosecutor v Deronjic* (Sentencing Judgment) IT-02-61-S (30 March 2004) para 257.

482 Office of the High Representative, 'High Representative Condemns Srebrenica Report' (OHR Press Release, 3 September 2002).

and stability' in the region.[483] He had done so whilst having 'full knowledge of the attacks perpetrated', according to the Prosecutor. While the charges under Common Purpose Liability were not confirmed and the Chamber established that 'press releases explicitly denying accusations of crimes levelled against the FDLR remain per se neutral,' it added that this neutrality is lost when 'it is demonstrated (i) that the [s]uspect knew that he was denying the truth; and (ii) that his denial of the truth was done in furtherance of an FDLR policy.'[484] Thus, although Mbarushimana's conduct was considered to have insufficiently contributed to the commission of the FDLR's crimes – and the charges were not confirmed – the Chamber did offer a two-tier test to potentially apply accessorial modes of liability to denialist disinformation.[485]

A final tension emerges between countering incitement and addressing denial as forms of discriminatory disinformation. While the incitement framework emphasises how manipulative techniques influence audiences, research on denialism demonstrates how perpetrators exploit this emphasis to deny responsibility, claiming they were merely following orders or were themselves manipulated. The defence strategies emerged during the Nuremberg trials and in the aftermath of the Rwandan genocide.[486] Interpretative denial as a strategy to evade responsibility may be inadvertently strengthened by overemphasising incitement's role in the commission of genocide. Legal frameworks must balance recognising disinformation's exploitation of cognitive and psychological weaknesses while affirming that exposure to such influences does not absolve individual culpability.

Beyond tension with international standards on protected speech, regulating denialist disinformation has caused diplomatic fallouts and inter-State accusations of intervention. When France in 2001 recognised the Armenian genocide and subsequently criminalised its denial in 2016, the Turkish government and press named these actions 'a great injustice', 'a total lack of respect for Turkey' and an 'intentional, malicious, unjust and illegal attempt' to change Turkey's history.[487] The Turkish ambassador to Paris was recalled and Türkiye issued a stream of threats of retaliation measures, after which France '[braced] itself for a political and economic backlash.'[488] While the threats never materialised,

---

483 *Prosecutor v Mbarushimana* (Decision on the Confirmation of Charges) ICC-01/04-01/10 (16 December 2011) paras 8, 10.

484 *Ibid.* paras 305, 213.

485 Mattias Holvoet, 'Disinformation in the Context of Mass Atrocity' (2022) 20 Journal of International Criminal Justice 1, 237-242.

486 Bilali, Iqbal and Freel (n 445) 286; Israel Charny (n 461) 15-39; S Buckley Zistel, 'Remembering to Forget: Chosen Amnesia as a Strategy for Local Coexistence in Post-Genocide Rwanda' (2006) 72 Africa 2, 131-150.

487 RFI, 'Turkey Prepares Retaliation Against France over Armenian Genocide Law' RFI (24 January 2012).

488 RFI, 'Marking 20 Years Since France "Upheld the Truth" and Recognised the Armenian Genocide' RFI (18 January 2017); NBC News, 'French bid to outlaw genocide denial outrages Turkey' NBC News (22 December 2011).

the incident illustrated how addressing denialist disinformation can strain diplomatic relations and potentially implicate non-intervention principles. The latter certainly applies when States intentionally instrumentalise denialist disinformation in the form of historical revisionism as a tool of foreign interference.[489]

In sum, genocide denial represents a distinct form of discriminatory disinformation that operates at the intersection of historical truth, collective memory, and legal accountability. International law does not comprehensively address denialist disinformation but sporadically references it in the balance of competing values of free expression, victim protection, and violence prevention.[490] The intent-harm paradigm offers the most coherent and viable approach for distinguishing between legitimate historical inquiry and harmful denial, particularly when integrated with existing hate speech prohibitions. Tangible application lies in restricting aggravated denialist disinformation under the rubric of unlawful hate speech. As with other forms of discriminatory disinformation examined, the effectiveness of legal responses depends not only on prohibition mechanisms but on understanding the unique psychological and social mechanisms through which denial perpetuates harm. By conceptualising genocide denial as discriminatory disinformation rather than merely contested speech, international law can better address its role in both obscuring past atrocities and enabling future ones.

5.5.4 False Allegations of Genocide

Where genocide denial represents disinformation about past atrocities, false allegations of genocide equally facilitate future harm or evade responsibility for actual misconduct. 'False allegations' comprise claims that genocide is ongoing or is imminent, without factual evidence. When created, produced and disseminated with the intent to cause harm, such allegations qualify as disinformation.[491] Their legal implications have received limited international attention despite growing significance in both inter-State relations and human rights contexts. Those who instrumentalise this type of disinformation exploit the rhetorical power of the term 'genocide'.[492] States are, for example, more

489 Christina Arribas *et al.*, 'Information Manipulation and Historical Revisionism: Russian Disinformation and Foreign Interference through Manipulated History-Based Narratives' (2023) 3 Open Research Europe 121.

490 E.g. Milanovic (n 447) 57-62 on holocaust denial as a violation of the right to private life.

491 These bad faith claims must be clearly distinguished from good faith disagreement about ongoing situations or preliminary investigations which lack disinformation's characteristic malicious intent.

492 William A Schabas, 'Atrocity Crimes (Genocide, Crimes against Humanity and War Crimes)' in William A Schabas (ed), *The Cambridge Companian to International Criminal Law* (Cambridge University Press 2015) 207.

inclined to support or approve international intervention when under the impression that genocide is ongoing or imminent, and individuals demonstrate greater willingness to undertake violent action against perceived threats of (imminent) genocide.[493]

False genocide allegations primarily operate through two mechanisms: as justifications for State violations of international law and as a catalyst for ethnically or racially motivated violence, often as a tool to divert attention from actual State or individual misconduct. Both tiers exploit the term's emotional and legal weight, but with different intended targets and outcomes. These allegations consistently appear as narratives within broader disinformation rather than as isolated claims. In inter-State relations, they surfaced in Russia's extensive disinformation campaign accusing Ukraine of committing genocide against Russian-speaking minorities living in the Eastern parts of the country (5.5.4.1). Here, the allegations function as a pretext for international law violations, particularly armed intervention.[494] The second manifestation finds clear illustration in "white genocide" or "great replacement" conspiracy theories, which serve to mobilise violence against minority groups through fabricated existential threats (5.5.4.2). While different in scope and actors, they both represent deliberate manipulation of genocide's legal concept to cause specific harms.

5.5.4.1 False Allegations in Interstate Relations: Ukraine v. Russia

The 'Nazi-genocide-Russophobia' frame for Ukraine traces back to the 1990's,[495] but gained unprecedented legal significance with the instituting of proceedings against the Russian Federation by Ukraine before the ICJ in 2022. This case constitutes a definitive illustration of disinformation's harm: Russia instrumentalised such false allegations as justification for its invasion, and Russia's official Investigative Committee even presented "evidence" that the crimes had taken place.[496]

---

493 Alexandra A Miller, 'From the International Criminal Tribunal for Rwanda to the International Criminal Court: Expanding the Definition of Genocide to Include Rape' (2003) 108 Dickinson Law Review 1, 362 in Jens D Ohlin, '#Genocide: Atrocity as Pretext and Disinformation' (2023) 63 Virginia Journal of International Law 2, 104; Mai-Linh K Hong, 'A Genocide by Any Other Name: Language, Law, and the Response to Darfur' (2008) 49 Virginia Journal of International Law 235, 238.

494 Waseem Ahmad Qureshi, 'Lawfare: The Weaponisation of International Law' (2019) 41 Houston Journal of International Law 1, 39-85.

495 Egbert Fortuin, '"Ukraine Commit Genocide on Russians: The Term "Genocide" in Russian Propaganda' (2022) 46 Russian Linguistics 313-347, 323-324.

496 Investigative Committee of the Russian Federation, "Russian Investigative Committee: All Persons Who Committed Crimes on the Territory of Donbas Will Be Identified and Held Accountable" (19 March 2022) (in Russian) Accessed 28 March 2024; Ministry of Foreign Affairs of the Russian Federation, 'Terrorist Crimes Committed by the Kiev Regime

The case's significance, however, extends beyond evidentiary questions to the legal status of disinformation. Many States intervening in the subsequent proceedings supported the position that 'non-violation complaints' without merit fall within the scope of Article IX of the Genocide Convention and thus the Court's jurisdiction.[497] They argue that false accusations of genocide may qualify as a dispute between the State parties, as it relates to the interpretation and/or fulfilment of the Convention.[498] As Ohlin notes, 'Russia, in *applying* the Convention's norms to the facts on the ground regarding Ukraine, falsely asserted that Ukraine has committed a genocide there, and Ukraine disagreed, thus generating a legal dispute regarding the application of the Convention.'[499] Beyond this textual interpretation, intervening States emphasise 'that the ICJ has the jurisdiction to declare the absence of genocide linked to the breach of the good faith obligations under the Convention as resulting in the abusive interpretation of the Convention's provisions.'[500] The Court's rejection of Russia's argument that 'the jurisdiction of the Court does not extend to allegations of violations' may not only be interpreted as an expansion of its jurisdiction regarding genocide,[501] but creates space for cases involving allegations without actual harm.

While the Genocide Convention provides the obvious legal instrument for this case, the often-overlooked 1936 Broadcasting Convention equally contains relevant provisions; a joint reading of the obligations to refrain from any form of incitement that may lead to war (Article 2) and the prohibition on incorrect statements which are likely to harm good international understanding (Article 3),[502] present a similar 'duty to engage in truthful and good-faith speech.'[503] Russia's false genocide allegations – even independent of their use as a pretext for the use of force – constitute bad faith interpretation of such

(12 September 2024) Report of the Ministry of Foreign Affairs of the Russian Federation (in English) Accessed 9 October 2024.

497 Republic of Latvia, Declaration of Intervention Pursuant to Article 63 of the Statute, Allegations of Genocide under the Convention on the Prevention and Punishment of the Crime of Genocide (Ukraine v Russian Federation) paras 40-44; ICJ UK Declaration, paras 32-33; ICJ Denmark Declaration, para 23; ICJ Estonia, paras 31-32; ICJ Ireland Declaration, paras 23-25; ICJ Finland Declaration, paras 29-33; ICJ Romania Declaration, paras 27-30; ICJ Portugal Declaration, para 27; ICJ Austria Declaration, paras 21-23; ICJ Greece Declaration, para 37; ICJ Australia Declaration, paras 34-38; ICJ Croatia Declaration, paras 26-31.

498 Iryna Marchuk and Aloka Wanigasuriya, 'Beyond the False Claim of Genocide: Preliminary Reflections on Ukraine's Prospects in Its Pursuit of Justice at the ICJ' (2023) 25 Journal of Genocide Research 3-4, 263-264.

499 Ohlin (n 493) 136; Allegations of Genocide under the Convention on the Prevention and Punishment of the Crime of Genocide (Ukraine v Russian Federation) (Application Instituting Proceedings) [2022] ICJ General List No 182, para 21.

500 Marchuk and Wanigasuriya (n 498) 264.

501 Ohlin (n 493) 137.

502 International Convention concerning the Use of Broadcasting in the Cause of Peace (adopted 23 September 1936, entered into force 2 April 1938) 186 LNTS 301, Articles 2 and 3.

503 Ohlin (n 493) 143-144.

obligations. However, given this Convention's dormancy and Russia's reservation to Article 7 of the Convention, this argument carries limited weight in the present case.

The implied obligation in the Genocide Convention to not 'falsely accuse another State of genocide' suggests, according to Ohlin, that the Convention 'includes a background norm of truthfulness regarding allegations of genocide.'[504] While this recognition connects to Russia's use of force against Ukraine, the Court's position does not contain *a priori* limitations to false allegations followed by the use of force, but appears to apply to questions on treaty application or fulfilment generally. If future developments support the existence of a general background norm of truthfulness and good faith in international relations as a ground for asserting jurisdiction,[505] this would position the ICJ as a key arbiter in countering disinformation. Although normatively desirable, it would redefine dispute resolution mechanisms under existing Conventions, which States may not eagerly agree with.

5.5.4.2 "White Genocide" Conspiracy Theories as Discriminatory Disinformation

False genocide allegations by non-State actors, particularly by extremist and terrorist groups are dangerous manifestation of disinformation. "White genocide" conspiracy theories and the fictional concept of a 'Great Replacement,'[506] claiming fabricated schemes to destroy the "white human race", illustrate this.[507] This political myth has gained traction among white nationalists across the United States, South Africa, Australia, New Zealand and Europe.[508] Its core narratives – that white populations face existential replacement – exist alongside racist speech promoting hatred against minorities. The past decade has witnessed multiple violent attacks motivated by these conspiracy theories. The 2019 Christchurch massacre of 51 Muslims in 2019 was explicitly linked to "white genocide" ideology, with the perpetrator's manifesto framing his actions as revenge for the genocide of white Europeans and ISIS terrorist attacks in Europe.[509] Similarly, the 2018 Pittsburgh synagogue shooting that killed 11 people was motivated by the claim that 'Jews

504 *Ibid.* 139.

505 *Ibid.* 143; Andrew Sanger, 'False Claims of Genocide Have Real Effects: ICJ Indicates Provisional Measures in Ukraine's Proceedings Against Russia' (2022) 81 Cambridge Law Journal 1, 217-221.

506 Renaud Camus, *Le Grand Replacement: Introduction au Remplacisme Global* (La Nouvelle Libraire 2011).

507 Mark Davis, 'Violence as Method the "White Replacement", "White Genocide", and "Eurabia" Conspiracy Theories and the Biopolitics of Networked Violence' (2024) Ethnic and Racial Studies, 6-7.

508 Francesco Farinello, 'Conspiracy Theories and Right-Wing Extremism – Insights and Recommendations for P/CVE' (Publications Office of the European Union 2021) 11.

509 *Ibid.* 11; Davis (n 503) 2; Gabriele Cosentino, *Social Media and the Post-Truth World Order: The Global Dynamics of* Disinformation (Springer International Publishing AG 2020) 77-79.

were committing a genocide on his people.'[510] Both attacks inspired further racially motivated violence, driven by similar motives of white supremacy and fears of 'complete racial and cultural replacement.'[511]

These false allegations of genocide predominantly qualify as unlawful hate speech and prohibited terrorist content. They clearly create 'to an atmosphere in which acts of racism, including acts of violence, are more likely to occur […]' under ICERD standards.[512] Constituting a clear example of propaganda containing 'ideas based on racial superiority or hatred', their proliferation is irreconcilable with meaningful implementation of the obligation to prohibit such racist speech.[513] These conspiracy theories, however, also assume an inter-State dimension when incorporated into State-operated or State-sponsored disinformation campaigns. Loefflad points out that the dissemination of "white genocide" conspiracies by Russia has become part of its broader orchestrated assault on the international order, particularly targeting stability and friendly relations.[514] Instrumentalising such allegations to promote racial hatred with the intent to 'aggravate the conflict between minorities and the rest of the population' in another State,[515] qualifies as disinformation interfering in the internal affairs of another State in contradiction of non-intervention obligations.[516]

False allegations of genocide with malicious intent represent a form of disinformation that can have far-reaching consequences for inter-State relations and internal stability and peaceful coexistence. While the narratives are similar, the harm intended runs in different directions: States and non-State actors may weaponise genocide terminology to justify violations of international norms, to divert attention from their own misconduct, and to enable violence. Depending on language, the group targeted, and the circumstances, false allegations of genocide may amount to advocacy for hatred, potentially rising to the level of incitement to violence. When propagated by States – either as pretext for international law violations or to influence foreign domestic politics – the applicable framework is more obfuscated, falling back on general international law frameworks, including non-intervention. Pending the ICJ's rulings in the

510 Lois Becket, 'More than 175 Killed Worldwide in Last Eight Years in White Nationalist-Linked Attacks' The Guardian (4 August 2019) Accessed 22 February 2025.

511 Davis (n 507) for an overview of contemporary events.

512 *The Jewish Community of Oslo and Others v Norway* (n 200) para 7.3.

513 ICCPR (n 87) Article 20(1); ICERD General Recommendation 35 (n 89); *The Jewish Community of Oslo and Others v Norway* (n 199) para 10.4; *JRT and the WG Party v Canada* (n 97) para 8(b); UNCHR 'General Comment 11: Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred' (29 July 1983) UN Doc HRI/GEN/1/Rev.9 para 2.

514 Eric Loefflad, 'International Law for a Time of Monsters: 'White Genocide', The Limits of Liberal Legalism, and the Reclamation of Utopia' (2024) 35 Law and Critique, 191-212, 197; William J Aceves, 'Virtual Hatred: How Russia Tried to Start a Race War in the United States' (2019) 24 Michigan Journal of Race & Law 177.

515 *Ibid.* 231.

516 Section '2.4 The Principle of Non-Intervention'.

Ukraine v. Russian Federation case, the incompatibility of false allegations of genocide with the Genocide Convention remains debated. While Russia's disinformation operations are undoubtedly incompatible with the 1936 Broadcasting Convention, this treaty's practical influence has been minimal since its adoption. Beyond specific legal obligations, the proliferation of false genocide allegations undermines the international community's commitment to prevent and punish genocide: malicious appropriation of 'genocide' diverts attention from actual genocides and intensifies the problem of the politisation of the term.[517]

5.5.5 Interim Conclusion

The interaction between discriminatory disinformation and genocide manifests in three distinct dimensions: disinformation that incites genocide and creates an environment conducive to incitement, disinformation that denies genocide occurred, and disinformation that falsely alleges ongoing genocide. International law provides regulatory guidance but also shows significant gaps. As Wilson argues, we are in the early stages of formulating a coherent doctrine on prosecuting speech crimes.[518] Several conclusions nevertheless emerged from this analysis. First, the overlap between discriminatory disinformation and incitement to genocide provides a clear pathway for legal intervention when it meets the requirement of being public, direct, and accompanied by genocidal intent. Second, while genocide denial often escapes direct prohibition, denialist disinformation targeting protected groups falls within the parameters of unlawful hate speech, with the intent-harm approach providing the most coherent strategy for distinguishing between legitimate historical and legal inquiries and harmful denial. Third, false allegations of genocide increasingly serve as tools to divert attention from wrongful behaviour, as pretexts for international law violations or as a catalyst for racially motivated violence. Despite their impact, they are only sporadically and indirectly addressed in existing legal doctrines.

5.6 Conclusion

This chapter has identified discriminatory disinformation as a multifaceted phenomenon operating at the intersection of disinformation and discriminatory

517 Elizabeth Whatcott, 'Compilation of Countries' Statements Calling Russian Actions in Ukraine "Genocide"' (Just Security, 20 May 2022); Michelle F Ringrose, 'The Politization of the Genocide Label: Genocide Rhetoric in the UN Security Council' (2020) 14 Genocide Studies and Prevention 1, 124-142.

518 Wilson (n 54) 11.

speech. The analysis revealed substantial convergence between hate speech, incitement and disinformation. Discriminatory disinformation can be identified by centralising the object of the message targeting group identity, the objective of the author ranging from influencing attitudes to inciting violence, and the contextual information environment that enables and amplifies its proliferation. Discriminatory disinformation exploits social inequalities while systematically targeting vulnerable groups and appealing to particularly susceptible audiences.

International human rights law provides the most suited framework for addressing discriminatory disinformation through existing prohibitions on incitement to discrimination, hostility, and violence. The contextual emphasis – focusing on content and presentation, socio-political climate, and speaker-audience dynamics – offers adaptable parameters to evaluate developing threats as unlawful hate speech. However, challenges persist regarding intent standards, proximity requirements for harm, and the proper understanding of 'hatred' and 'hostility'. Here the distinction between 'imminence' and 'likelihood' – while largely neglected in scholarship and jurisprudence – proves significant in capturing discriminatory disinformation's gradual, cumulative effects opposite immediate incitement.

Within international criminal law, discriminatory disinformation meeting the stringent requirements of being public, direct, and accompanied by genocidal intent falls within the prohibition on incitement to genocide. Yet digital information environments challenge traditional interpretations, particularly regarding public/private distinctions and the temporal dimensions of incitement. While not unique to this regulatory framework, these challenges are particularly potent considering the 'rigid and demanding' threshold for incitement to commit genocide. The analysis likewise identified how discriminatory disinformation manifests in genocide denial and false allegations of genocide – forms that may not constitute direct incitement but nevertheless contribute to environments conducive to mass violence or other forms of discrimination. These dimensions reveal how disinformation serves both to enable and obscure atrocities in a dangerous cyclical pattern.

Overall, this application exposed a disconnect. While international law contains substantive norms applicable to this phenomenon, practical effectiveness is constrained by a lack of adaptation to technological developments, evolving speaker-audience relationships, and the manipulative nature of contemporary information operations. As the threat expands, the frameworks' protective reach seems to contract. Technological amplification, algorithmic targeting, and coordinated anonymous dissemination create implementation challenges traditional frameworks struggle to capture, let alone meaningfully respond. Nevertheless, existing legal architecture, when applied with attention to disinformation's sociotechnological nature, retains significant potential for curbing this harmful phenomenon.