Speech Crimes in the Digital Era

The Peace and Justice Initiative Guidelines

Edited by **Matthew Gillett, Vanessa Topp *and* Richard Ashby Wilson**



BRILL | NIJHOFF

Speech Crimes in the Digital Era

Speech Crimes in the Digital Era

The Peace and Justice Initiative Guidelines

Edited by

Matthew Gillett
Vanessa Topp
Richard Ashby Wilson

With

Juan Manuel Klein
Damla Yasemin Ipek
Katharina Braun
Manuel J. Ventura



BRILL | NIJHOFF

This publication was supported by the University of Essex's open access fund.

This work was supported by The Peace and Justice Initiative.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. https://creativecommons.org/licenses/by-nc-nd/4.0/

This book is available as an open access publication via www.degruyterbrill.com.

ISBN 978-90-04-78025-5
ISBN 978-90-04-78026-2 (PDF)
DOI https://doi.org/10.1163/9789004780262

Library of Congress Control Number: 2026949699

© 2027 the author(s)/editor(s), published by Koninklijke Brill BV, Plantijnstraat 2, 2321 JC Leiden, The Netherlands

Brill Nijhoff and Koninklijke Brill BV are part of De Gruyter Brill.
www.degruyterbrill.com

Questions about General Product Safety Regulation: productsafety@degruyterbrill.com

Cover illustration: Magnus Zeller, *The Orator*, circa 1920, Los Angeles County Museum of Art, purchased with funds provided by Charles K. Feldman, photo © Museum Associates/LACMA.

Contents

Foreword by Judge Howard Morrison IX
Preface XI
Executive Summary XIII
Key Guidelines XX

Introduction: Defining and Conceptualising Speech Acts in the Digital Era 1
1 Speech Act Theory 3

1 Interpreting and Analysing Speech Acts 5
1 Linguistic Justice: the Interpretation of Speech in Context 7
1.1 Case Vignette: Ntaganda at the ICC 8
1.2 Case Vignette: Krstić and Popović et al. at the ICTY 11
1.3 Case Vignette: Interpretation at the Special Court for Sierra Leone 12
1.4 Case Vignette: the Importance of Socio-Linguistic Experts for Fair Trials – a Lesson from Western Australia 14
2 Forensic and Legal Analysis of Speech Acts 16
2.1 Orders as Speech Crimes 17
2.2 Case Vignette: Ntaganda ICC Trial Judgement at the ICC 18
2.3 Case Vignette: Abd-al-Rahman ICC Trial Judgement at the ICC 21
2.4 Case Vignette: Karadžić IRMCT Appeals Judgement at the ICTY 24
2.5 Case Vignette: Mladić IRMCT Appeals Judgement at the ICTY 25

2 Disinformation and Speech Crimes 28
1 Conceptualising Disinformation in International Trials 31
2 Disinformation and Propaganda as Addressed by International Courts 32
3 Case Vignettes on Disinformation Campaigns and Atrocity Crimes 34
3.1 Case Vignette: Fritzsche at the IMT at Nuremburg, 1945 34
3.2 Case Vignette: Julius Streicher at the Nuremberg Tribunal 38
3.3 Case Vignette: Gvero at the ICTY 39

3.4 Case Vignette: Mladić at the ICTY 44
3.5 Case Vignette: Karadžić at the ICTY 49
3.6 Case Vignette: Mbarushimana Confirmation of Charges at the ICC (Pre-Trial Chamber) 54
3.7 Case Vignette: Ruto and Sang Confirmation of Charges at the ICC 59
4 Analysis – Disinformation in International Criminal Law 62
4.1 Disinformation as a Contribution to a Crime (Modes of Liability) 64
4.2 *Actus reus* (the Physical Element) 65
4.3 *Mens rea* (the Mental Element) 65
5 Disinformation in the Form of Genocide Denial and Its Treatment in Rwanda and Germany 68
5.1 Genocide Denial in Rwanda 68
5.2 Memory Laws in Germany 70

3 Social Media and Speech Crimes 74
1 A Pre-digital Era Case 75
1.1 Case Vignette: Nahimana et al. (the Media Case) at the ICTR 75
2 Incitement to Genocide and Other Atrocity Crimes on Social Media 77
2.1 Case Vignette: Rohingya in Myanmar 78
2.2 Case Vignette: Russia and Ukraine 81
2.3 Case Vignette: Israel and the Occupied Palestinian Territory (oPt) 86
3 Pathways to Liability for Social Media-Driven International Crimes 94
3.1 Corporate Criminal Liability for International Crimes 94
3.2 Social Media Algorithms and Liability 100
3.3 Liability of Natural Persons for Social Media Corporations 103
3.4 Liability of Social Media Executives 108
4 Establishing Jurisdiction over Speech Crimes Disseminated through the Internet, including on Social Media Platforms 116
5 The Approach Taken by Domestic Courts to Criminal Activity Online 119
6 At the ICC 122
7 Social Media Speech Regulation 125
7.1 A Contextual Analysis of Online Speech 131
8 Recommendation: a United Nations Digital Code of Conduct 133

4 Gathering and Submitting Evidence of Speech Crimes 138
1 Types of Evidence of Speech Crimes on Social Media 139
2 Obtaining, Archiving and Preserving Potential Evidence 141
3 Evaluating Potential Evidence 142
4 Verification of Potential Evidence of Speech Crimes 144
5 Presenting Social Media Speech Crimes Evidence 147
6 Fair Trial Concerns 151

5 Universal Jurisdiction Cases and Speech Crimes 153
1 Domestic Prosecution of Speech Charged as a Crime 154
2 Domestic Prosecution of Direct and Public Incitement to Genocide 155
3 Domestic Prosecution of War Crimes and Crimes against Humanity through Speech Acts 156
4 Domestic Prosecution of Speech as a Mode of Liability 157
4.1 Belgium: Perpetration of War Crimes and Genocide 158
4.2 Germany: Aiding and Abetting War Crimes 158
5 Cross-Cutting Issues in Universal Jurisdiction Cases 160

6 Direct and Public Incitement to Aggression, Crimes against Humanity, and War Crimes 162

7 Key Guidelines for the Assessment of Speech Crimes under International Criminal Law 168
1 Updating the Risk Assessment Model for Speech Crimes 170
1.1 Risk Assessment Model for Speech Crimes in the Digital Age 170

8 Concluding Remarks 173

Index 183

Foreword by Judge Howard Morrison

Although speech crimes in international law have been around for longer than one might think, this threat has gained great impetus in the era of digital communications, social media, and the rise of artificial intelligence generated discourse. All too often, criminal speech continues to be found in the announcements of certain agitators and political commentators.

There are striking examples of speech crimes in the propaganda from the Nazi party in the 1930s, which can also be seen in messaging disseminated by the Kremlin from the 1930s right up until today. We have political leaders who exaggerate and vilify certain actions and people for their own political partisan ends, all while the distinction between free speech and a speech crime has become worryingly indistinct. There are those in power today who pay little regard to international law and who are actively hostile to the institutions that promote accountability. We live in disturbing times.

I first became involved in identifying speech crime in international law when I was defending at the Rwanda tribunal. *Radio Mille Collines* was propagating violent anti-Tutsi sentiment, calling Tutsis 'cockroaches', amongst other names, and it was all designed to do what it did, namely, to promote the dreadful killings. Hate and crime speech are in practice virtually indistinguishable and we face a very challenging task in the world of fake news and profound exaggeration that is utilised not only by individuals with prejudicial grievances but often by state media, by many unethical political parties and even those in positions of great personal power.

It therefore is more critical than ever for international lawyers and international observers to understand the ramifications and the dangers that lie ahead. This erudite initiative, representing as it does a significant new contribution to the literature and advancement from the Hartford Guidelines, recognises the very fundamental technological developments and the increasingly complex role of speech in the commission, facilitation, and prosecution of international crimes. It thus assists in advancing accountability and setting out templates for lawyers and others involved to uncover and promote a sense of justice in this complex landscape.

The fundamental right of free speech is, I fear, being instrumentalised in the name of partisan political views, some extremely violent and dangerous, in order to prevent or at least divert accountability when inflammatory speech incites or contributes to international crimes, even the crimes of aggression and genocide. A balance has to be drawn between the preservation of the genuinely fundamental and democratic principles that underpin free speech,

while not allowing those principles to be abrogated by those who will use or misuse speech to promote violent political or criminal agendas.

This publication will assist those of us who find the complexities of technological advance and artificial intelligence hard to comprehend and enable a wide spectrum of people to see where the dangers lie. History shows us that speech in the form of propaganda, or the promotion of social division, frequently precedes and contributes to the commission of mass atrocities. This has been significantly intensified through the use of digital and AI-generated communications where reality can be distorted and information fabricated with great sophistication so that it is extraordinarily difficult to distinguish truth from lies. In my judgement, incitement to commit crimes has been a neglected tool in the armoury of international prosecutors, but that will change when people fully realise the dangers that we face. This publication endorses a two-stage analytical approach, it enables the identification of disinformation and illustrates the broad scope of social media being misused for partisan ends. This will lead, when properly used, to more sophisticated indictments and that in turn will require a deeper and greater understanding by prosecuting lawyers and trial and appellate judges. I do not ignore defence lawyers and it is going to become important for those defending hate and speech crimes to understand the full parameters of the cases with which they are dealing.

These advances can be utilised in existing and future tribunals, universal jurisdiction prosecutions and the domestic prosecution of international crimes. It is going to be important to look at prosecutions beyond individual criminal responsibility into the realm of corporate responsibility when the relevant gravity threshold is established.

I agree at the international level States should be required to affirm their obligations to prevent speech crimes, which can be done by adopting an amendment to Article 25(3)(e) of the Rome Statute to expand it to include the incitement of any Rome Statute crimes. I thank the authors for their careful and erudite work and commend this publication to all of those engaged in pursuing accountability for this very fundamental issue in international criminal law. (Former Judge of the International Criminal Court and the International Criminal Tribunal for the Former Yugoslavia).

Preface

These Peace and Justice Initiative Guidelines build on the Hartford Guidelines on Speech Crimes in International Criminal Law.[1] The contents are entirely new with one exception – the renewed call to amend International Criminal Court Statute Article 25(3)(e) – due to the continued prevalence of speech crimes that have increased with the rapid advance and adoption of new digital technologies. Indeed, much has changed since the Hartford Guidelines were published, including the ubiquity of digital platforms and the increasing introduction of digital evidence of speech crimes in international criminal proceedings. The world also experienced the outbreak of several major wars, in which the digital domain has become a new conflict front, and the Covid-19 pandemic, which brought digital technology to the forefront as businesses, individuals, and families embraced online meeting platforms and new social media enterprises to a greater degree than ever before. Equally, a number of important decisions and judgements applied international criminal law to issues relevant to speech crimes.

Against this backdrop, the need to update our thinking about international legal doctrine and procedure relating to speech crimes in the digital age is evident. In preparing these guidelines, the co-editors were fortunate to receive inputs from an outstanding set of academics and professionals, ranging from those in the early stages of their careers to those with extensive experience of international criminal law and related fields. The co-editors thank Katharina Braun, Juan Manuel Klein, Damla Yasemin Ipek, and Manuel Ventura* for their substantive contributions, as well as Predrag Dojčinović, Wibke Timmermann, Marija Ristić, Iva Vukušić, Elena Sherstoboeva, Sarah Zarmsky, Mathias Holvoet, Franziska Fluhr, Rachel Chambers, David Kaye, Elaine Gao, and Angela Li.

---

* The views here do not necessarily reflect those of the Australian Government.

1 In light of the growing ubiquity of social media and artificial intelligence, the Peace and Justice Initiative together with the University of Essex and Princeton University's Human Rights Initiative has produced these guidelines on speech crimes, to add a new volume of research and expand on the Hartford Guidelines. The Peace and Justice Initiative is a Dutch-registered non-governmental and non-profit organisation established to encourage accountability for atrocity crimes through judicial proceedings. It previously produced The Hartford Guidelines on Speech Crimes in International Criminal Law (2018) together with the University of Connecticut School of Law and Human Rights Institute, and it provided a submission to the International Law Commission for its consideration in formulating the proposed Crimes Against Humanity Convention.

The co-editors also reiterate their thanks to all involved in the production of the Hartford Guidelines, which provided the foundation on which these updated Peace and Justice Initiative Guidelines are built. This present project was fortunate to receive support from the Peace and Justice Initiative, the University of Essex, and Princeton University. Through these Guidelines, the co-editors seek to contribute to the prevention and mitigation of atrocity crimes. Addressing speech crimes through international criminal law is not a panacea, but it is one important facet of the struggle against the large-scale and severe repression of targeted populations around the world. Those who use words to ignite violence should face full accountability for the resulting harms, whatever the medium used to disseminate those words and whether enhanced and intensified by artificial intelligence or other digital technology. Harnessing the authority of international criminal law to assist in this endeavour can help deter such incitement and instigation. Yet the normative framework of international criminal law can benefit from certain doctrinal amendments, most notably through an amendment to Article 25(3)(e) of the Rome Statute, as proposed here. We encourage the widespread use of these Guidelines to assist with applying legal sanctions to serious speech crimes falling outside the fundamental protections of free expression and belief.

Executive Summary

1. This publication, titled the Speech Crimes in the Digital Era: The Peace and Justice Initiative Guidelines (hereafter, the 'PJI Guidelines'), constitutes an innovative new body of research, which expands upon the Hartford Guidelines on Speech Crimes in International Criminal Law (hereafter, the 'Hartford Guidelines'). It assesses international and domestic cases relevant to speech crimes adjudication in depth, and pays particular attention to analysing how international criminal law has handled new developments in digital technology and Artificial Intelligence (AI) and presents new ways of approaching speech crimes in the digital era. The PJI Guidelines examine six dimensions of speech crimes, ranging from the analytical to the normative, and propose reform of the Rome Statute of the International Criminal Court (Rome Statute) on the basis of this analysis. The 'Key Guidelines' distil the main points relevant to the practice of adjudicating speech crimes on the basis of international law principles in the current age.

2. The proposals contained in the Key Guidelines are consistent with the axiomatic human right to freedom of expression. As restated by the Human Rights Committee (HRC) in General Comment 34, any restrictions on freedom of expression must comply with the conditions of Article 19(3) of the International Covenant on Civil and Political Rights (ICCPR): the restrictions must be 'provided by law'; they may only be imposed for one of the grounds set out in subparagraphs (a) and (b) of paragraph 3; and they must conform to the strict tests of necessity and proportionality.[1] Given the nature of the crimes listed under the Rome Statute, namely war crimes, crimes against humanity, genocide and aggression, any speech inciting, assisting, encouraging or otherwise calling for or contributing to such crimes would in principle be open to criminalisation, depending on the context of the speech.

3. However, the form in which inciting speech is criminalised will determine the legality of the restriction on freedom of expression and must not be vague, overbroad, or retroactive. Speech is a prominent feature of leadership trials under international criminal law – as can be seen most starkly in the recently initiated case against former Philippines President Rodrigo Duterte.[2] In this context, ensuring a precise understanding of the potentially applicable speech

1 United Nations Human Rights Committee (HRC), 'General Comment No. 34, Article 19: Freedoms of opinion and expression' (102nd session Geneva, 11–29 July 2011) CCPR/C/GC/34 § 22.

2 See *Prosecutor v Rodrigo Roa Duterte*, 'Corrected Version of 'Prosecution's Pre-Confirmation Brief' ICC-01/21-01/25-201-Corr-Red2 13-02-2026 2/82 PT (13 February 2026).

crimes, referring to both substantive crimes and modes of liability, is critical. To this end, the following analysis is anchored on the twin precepts of imposing accountability and adhering to the Court's human rights foundations.

4. Whereas the PJI Guidelines are designed to be implemented in the practice and application of the law, they are founded on social research and legal scholarship as well as doctrinal legal analysis. The analysis draws on a broad selection of empirical data, particularly Judgements and decisions of the International Criminal Court (ICC), the International Criminal Tribunal for the former Yugoslavia (ICTY), the International Criminal Tribunal for Rwanda (ICTR), and other international and hybrid courts. They also incorporate the views of leading scholars, Government reports, United Nations (UN) instruments and issuances, and expert analyses.

5. Building on these sources, the PJI Guidelines highlight the fact that speech acts frequently precede and incite the occurrence of inter-group and/or large-scale kinetic violence. The risk of violence is heightened and extended geographically in the digital and artificial intelligence age, as inflammatory messages can be instantly broadcast or directly sent to recipients around the world and algorithmically amplified. They can also be manipulated, edited, and even artificially generated in the first place. With material conflicts regularly accompanied by parallel information wars online, the threat of direct and public incitement to atrocity crimes as well as the instigation of perpetrators who commit crimes are real and present dangers.

6. Against that backdrop, the PJI Guidelines make novel and important contributions to this field. First, they advocate a two-stage process for the evaluation of speech acts, in which the statements are interpreted in their context in order to establish their factual meaning and impact, followed by an analysis of the legal significance of the statements on the basis of the applicable normative framework. Second, the PJI Guidelines identify major factors critical for the assessment of speech acts in the digital and AI age, including preservation, amplification, and temporality. Third, they examine the notions of disinformation, social media, universal jurisdiction, and corporate responsibility in relation to speech crimes. Finally, as a matter of statutory reform, the PJI Guidelines recommend adjusting Article 25(3)(e) of the Rome Statute, which currently only concerns incitement to genocide, so that the amended version would prohibit 'intentionally, directly, and publicly inciting others to commit any of the crimes contained in the Statute, thereby significantly increasing the likelihood of their occurrence'. By criminalising all direct and public incitement of atrocity crimes, and not only the incitement of genocide, the international community can add an important preventive tool to facilitate accountability

and deterrence. Mechanisms for preventing mass atrocities, as observers have noted frequently, are underdeveloped in international criminal law.[3]

1 Chapter 1: Interpreting and Analysing Speech Acts

7. Accurate interpretation and contextual analysis are fundamental to investigating and adjudicating potentially criminal speech acts. The approach to linguistic analysis and interpreting speech in international trials, especially those concerned with a speech act, demonstrates that a two-stage analytical process should be followed, with the meaning of the words first established within their context and then the legal significance of the utterance(s) assessed through the applicable normative and legal framework.

8. To this end, the specific words used in an utterance must be interpreted in context and in view of the manner of their usage. This includes the relevant cultural and historical context and also requires taking into account both the military context and the broader social and political milieu that are contemporaneous with the speech act. Interpretation and contextualisation of utterances should be based on witness and documentary evidence, including from experts on the relevant cultural and linguistic context. It may also include testimony from insiders from the forces to whom the utterance was made as well as fact witnesses, or international actors involved in resolving or mediating the conflict. International and national criminal justice agencies stand to benefit from guidance on how to address contesting accounts regarding the meaning of legally significant cultural manifestations or context specific utterances. This is particularly important when said utterances may have contributed to the commission of international crimes and as such, their accurate and fair interpretation can play a key role at trial – especially in the case of languages of lower diffusion.

2 Chapter 2: Disinformation and Speech Crimes

9. Disinformation poses a unique and growing risk in the digital and AI era. Defined as the intentional dissemination of false information with the purpose

3 William Schabas, *Genocide in International Law: The Crime of Crimes* (2nd edn, Cambridge University Press 2009) 521.

to cause serious harm,[4] disinformation is increasingly utilised by perpetrator groups to further criminal enterprises. The dissemination of misleading and purposefully mendacious information can inflame tensions and sectarian animosities, heightening the risk of atrocity crimes being committed. In its most acute form, disinformation itself can constitute a speech crime. Disinformation not only has the potential to directly incite crimes, but it may also play a role as a contributing factor in the commission of crimes. For example, by distracting from an offence, disinformation can contribute to a climate of impunity, or it can mislead the international community as part of a criminal enterprise or conspiracy to commit international crimes. As such, temporally the dissemination of disinformation may perform different functions and fulfil the elements of different modes of liability at different points in relation to the commission of atrocity crimes including prior to (inciting), during (facilitating), or after (cover-up, a form of aiding-and-abetting). Any analysis of potential liability for disinformation as a purposeful act intended to manipulate or mislead its audience and support a larger plan to commit atrocity crimes should consider whether the suspect was aware of the crimes (whether planned or committed), whether the suspect deliberately disseminated untrue information, and whether the suspect did so to further criminality, as well as the impact that the deception had on the perpetration or perpetrators of the crime(s).

10. While disinformation may be disseminated through traditional forms of media, the growth of social media platforms in recent years has allowed for a frictionless and viral proliferation of disinformation to a wider audience – bringing increased attention to this issue and to its potential role in the commission of international crimes in recent or ongoing conflicts. The rapid breakthroughs in generative AI technologies further facilitate the creation and spread of disinformation as convincing and increasingly realistic content in the form of text, imagery, audio and video can be fabricated and disseminated at a rapid pace.[5] As these technologies continue to evolve, there is a lack of coordinated approach for moderating and responding to AI-generated content and disinformation on the part of international legal bodies. Indeed, while AI tools themselves may have potential to analyse and detect falsified content

4 Here the PJI Guidelines adopts the approach of the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression Irene Khan as defined in UN General Assembly (UNGA), Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 'Disinformation and freedom of opinion and expression' (13 April 2021) A/HRC/47/25 § 15.

5 Noémi Bontridder and Yves Poullet, 'The Role of Artificial Intelligence in Disinformation' (2021) 3 Data & Policy e32.

and disinformation, its use for content moderation is not without safeguarding challenges and human rights concerns linked to freedom of expression.

3 Chapter 3: Social Media and Speech Crimes

11. Social media platforms can play a dual role in the commission and prosecution of speech crimes. Social media has provided the vehicle for the proliferation of disinformation and hate speech, which can facilitate the rapid and wide distribution of violent and incendiary messages to a large audience. On the other hand, social media presents new avenues and possibilities for holding perpetrators of atrocity crimes to account through enhanced gathering, collecting and storing of evidence of crimes as they are taking place. While this evidence gathering allows remote access to previously inaccessible conflict situations and facilitates communication with victims, investigations also come with their own risks and challenges. This section details several recent case studies in which speech crimes were allegedly spread through the use of social media. The case studies are followed by a discussion on emerging accountability gaps, the risk of an environment of impunity around social media, the effect of the use of automated algorithmic models and recommender systems on liability, and potential avenues towards corporate responsibility for social media companies. Natural persons using social media corporate structures to commit speech crimes can be held individually accountable irrespective of the availability or unavailability of corporate responsibility mechanisms.

4 Chapter 4: Gathering and Submitting Evidence of Speech Crimes

12. With an increased focus on speech acts in international criminal litigation, gathering and submitting evidence of this type of liability is a critical topic. The overarching framework governing evidence at the ICC is well-established in relation to 'traditional' crimes. Yet this framework is underdeveloped with respect to speech crimes. Evidential standards may require considerable adjustment in the future as more cases feature speech components and the jurisprudence addresses issues of admissibility and weight. Evidence of speech crimes will often be extracted from online sources, which is a dynamic evidentiary area in need of clarification and the development of standard protocols. When it comes to speech acts, there are several evidentiary questions that will affect the probative value of the materials and the manner in which they are

presented, including the metadata, the analysis of the contents including by experts, assessments of tampering or deepfakes, chain of custody, and the best evidence rule. In this section, these guidelines explore the types of evidence most relevant for this purpose. They evaluate the means of identifying, archiving and preserving these forms of digital evidence and the methods by which adjudicative bodies might verify such materials.

5 Chapter 5: Universal Jurisdiction Cases and Speech Crimes

13. Universal jurisdiction, the practice by which states exercise jurisdiction before national courts over crimes committed outside of their territory and without the victim or the perpetrator being a national of the prosecuting state, is becoming an increasingly important avenue for the prosecution of international crimes. As such, universal jurisdiction implies a tension between state sovereignty and effective international prosecutions of serious crimes and is therefore typically reserved for grave offences in international law. In principle, speech crimes can be prosecuted domestically via universal jurisdiction as a stand-alone crime (such as direct and public incitement to genocide) or as a mode of liability (such as inducing crimes against humanity) in ways that are similar to international courts and tribunals. In practice, there have been a limited but growing number of universal jurisdiction cases featuring a speech crime element in recent years. From these cases, we note that prosecuting speech crimes under universal jurisdiction may face several challenges. Although domestic legislation on international crimes may be modelled after international law, the successful prosecution of speech acts may ultimately hinge on national investigative tools, laws, and doctrines at the national level. Furthermore, there is a potential risk of national authorities overlooking or underestimating the role of speech due to a lack of knowledge regarding the cultural context of crimes and the meaning of specific words and phrases as it pertains to this context.

6 Chapter 6: Direct and Public Incitement to Aggression, Crimes Against Humanity, and War Crimes

14. Currently, the preventative function of speech crimes remains underrealised. Specifically, Article 25(3) of the Rome Statute of the ICC positions direct and public incitement to commit genocide alongside modes of liability instead of as a stand-alone inchoate crime. By introducing ambiguity as to

whether incitement to genocide is an inchoate stand-alone crime or a contingent mode of liability that potentially requires the commission of genocide before a prosecution can be mounted it undermines the preventative value of incitement as an inchoate crime, as envisaged in the UN Genocide Convention of 1948. Although Articles 25(3)(b) to (d) can technically apply to crimes that are attempted or committed, that requires an additional element beyond purely inchoate liability for incitement. Conversely, Article 25(3)(e) does not require that the actual crime be committed or even attempted. To strengthen this potential preventative function of international criminal law, Article 23(3)(e) of the Rome Statute should be expanded to include inciting any of the Rome Statute crimes, as formulated in the suggested amendment below.

15. The suggested amendment to Article 25(3)(e) of the Rome Statute reads as follows (with explanatory discussion set out in Substantive Chapter 6 below):

> Intentionally, directly, and publicly incites others to commit any of the crimes contained in the Statute, thereby significantly increasing the likelihood of their commission. For the purpose of this provision, it is not necessary that the incited crime(s) be committed or attempted.

Key Guidelines

16. We propose 10 Key Guidelines for the assessment of speech crimes. These Key Guidelines are designed as succinct reference points, focusing on the most important facets of assessing speech crimes under international criminal law. They do not purport to cover all issues relevant to speech crimes and should be interpreted and applied according to the detailed discussion set out in this entire document, along with the Hartford Guidelines. Although the Key Guidelines are largely analytical, operational, and doctrinal, they also include normative proposals for statutory reform aimed to expanding the preventive utility of international criminal law by amending Article 25(3)(e) of the Rome Statute to cover all crimes under the Statute and not just incitement to genocide. The Key Guidelines should also be read against the background of rapidly evolving technology and increasing diffusion of social media and AI, necessitating a flexible approach to their application to emerging and new circumstances, platforms, and technologies. Finally, the Key Guidelines are not formulated as legal rules per se, though they constitute important principles for the generation of a more regulatory instrument focused specifically on the investigation and adjudication of speech crimes.

1. Speech acts that may incur criminal liability should be evaluated in the context of their usage. Analysing speech acts can be facilitated by conducting a two-step process: first, the linguistic interpretation of the meaning of the statement(s) made within the relevant context; and, second, the legal assessment of the significance of the statement(s), and particularly any contribution that the statement(s) made to the alleged crimes.[1]
2. Governments and digital companies removing unlawful content should ensure that evidence of speech crimes is preserved. Actors researching and investigating speech crimes linked to social media should verify content incorporating up-to-date technological tools and recording their investigative steps.
3. Analysing alleged speech crimes linked to social media requires taking account of the number of views and amplification of the speech act and the speaker's socio-political role, position, authority, and reach.

---

1 These consider the 14 factors identified in the Hartford Guidelines relating to the status of the speaker, the content of the message and the context of the speech act) form part of the second step of this process. See Richard Ashby Wilson and Matthew Gillett, *The Hartford Guidelines on Speech Crimes in International Criminal Law* (Peace and Justice Initiative 2018) § 320.

4. Disinformation is multi-faceted as it can directly contribute to crimes and can also distract from offences and thereby contribute to a climate of impunity or mislead the international community. Temporally, disinformation prior to (incitement), during (facilitating), or after (cover-up) crimes should be investigated.
5. Assessments of alleged disinformation as a contribution to atrocity crimes should address whether (i) the suspect had knowledge of the crimes (whether planned or committed); (ii) the suspect intentionally and knowingly disseminated untrue information; and (iii) the suspect did so to further the criminality.
6. Speech crimes charged under international criminal law must meet a gravity threshold to qualify for prosecution. If it is charged as a mode of liability, the speech act must satisfy the requisite contribution threshold under that mode to the crime (or crimes) which meets the gravity threshold. The speech act should be more than a remote occurrence in the sequence of events leading up to the crime. If the speech itself is alleged to constitute the crime, it must be of sufficient seriousness to meet the requirements of that crime. In applying this threshold, courts should take into account the nature of the speech act, its repetition, its reach, and its impact.[2]
7. In universal jurisdiction cases in domestic systems involving speech acts, these recommendations should be incorporated into national laws, doctrines, and investigative methods. All cases prosecuted under universal jurisdiction should adhere to internationally recognised due process and human rights principles.
8. Corporate responsibility should be extended to speech crimes and should be implemented in line with these recommendations. Criminal responsibility should only be applied where the relevant gravity threshold is met. Other regulatory sanctions such as retractions, fines, restriction of licence, and dissolution can be considered for alleged speech crimes which do not meet the gravity threshold (and could also be imposed along with criminal responsibility where appropriate), though always bearing in mind the fundamental right of free expression. Natural persons using corporate structures to commit speech crimes should be held

2 This Key Guideline does not seek to add an element to any crime but rather highlights the relevance of the gravity admissibility requirement under the Rome Statute. See *Rome Statute of the International Criminal Court* (adopted 17 July 1998, entered into force 01 July 2002) 2187 UNTS 3, art 17(1)(d).

accountable irrespective of the availability of corporate responsibility mechanisms.

9. To strengthen the preventative function of international criminal law, States should affirm the obligation to prevent speech crimes. At the international level, this goal should be pursued by adopting an amendment to Article 25(3)(e) of the Rome Statute to include inciting any of the Rome Statute crimes, as formulated below. The new article confirms the inchoate nature of incitement by not requiring an atrocity crime to occur in order to prosecute its intentional, direct and public incitement:

> Intentionally, directly, and publicly incites others to commit any of the crimes contained in the Statute, thereby significantly increasing the likelihood of their occurrence. For the purpose of this provision, it is not necessary that the incited crime(s) be committed or attempted.

10. States, with the assistance of civil society, and other relevant entities should implement these recommendations. State Parties to the Rome Statute should consider the amendment to Article 25(3)(e) and propose it to the Assembly of State Parties for referral to the Working Group on Amendments. The drafters of these Guidelines stand ready to assist with the consideration and implementation of the proposed amendment.

INTRODUCTION

Defining and Conceptualising Speech Acts in the Digital Era

17. The overarching aim of these PJI Guidelines is to build on the foundation provided by the Hartford Guidelines, in light of the increasing ubiquity of digital technology and AI. The Hartford Guidelines provide a detailed survey of key international trials with speech elements and highlight the fact that speech crimes remain an unsettled area of international criminal law. They demonstrated that the uncertainty around the legal regulation of speech crimes means that its preventative potential is as yet unrealised. The Hartford Guidelines noted that 'there is confusion regarding the inchoate status of incitement to genocide, a lack of clarity about whether hate speech could constitute an element of persecution (a crime against humanity) in the Statute of the International Criminal Court, and general uncertainty regarding the evidence required to demonstrate causation for modes of liability such as instigation'.[1] The advent of the digital era and the growing use of AI has further undermined the preventive function of speech crimes in international criminal law. Moreover, procedures governing the admission, exclusion, and weighing of evidence need to be revisited in light of the growing prevalence of digital evidence and the increased use of AI to discover, organise, and analyse evidentiary material. These Guidelines seek to assist scholars, practitioners, State representatives, civil society and other relevant actors in confronting the threat of speech crimes in these technologically-driven times.

18. The challenges associated with creating an appropriate response framework to regulate inciting speech and subsequent human rights violations have taken on a new sense of urgency due to the growing ubiquity of social media – which has created a widespread platform for the commission of speech crimes. As recognised in the Hartford Guidelines, it is crucial to reiterate that the protective potential of sanctioning prohibited speech through international criminal law must be realised alongside attendant obligations to protect free speech enshrined in international human rights law.

19. As a new and complementary volume building on the Hartford Guidelines, the PJI Guidelines expand on some of the topics previously addressed and consider novel topics, such as the growing role of social media in conflict,

1 Wilson and Gillett (Key Guidelines, n 1) § 2.

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-002
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

in particular the potential for speech acts and disinformation disseminated through social media to contribute to the commission of international crimes. These Guidelines support this analysis on an empirical basis considering the role of speech crimes and disinformation for the specific case studies such as those concerning the Rohingya in Myanmar, Russia and Ukraine, and Israel and the occupied Palestinian territory (oPt). They also draw on theoretical literature and doctrinal discussions by experts in the field of speech crimes. Whilst a key focus on the PJI Guidelines is the digital era, there are several sections of the following analysis which address more traditional style evidentiary considerations, such as the meaning of orders in the *Ntaganda* case. Those analyses are important both in and of themselves, as markers for speech crimes assessment, and also for the foundational principles that they imply for digital media and AI technologies.

20. In each of these contexts, disinformation, hate speech and incitement have violated the human rights of targeted groups and have intentionally contributed to the commission of atrocity crimes by entrenching dangerous societal divisions, provoking violence, and mobilising groups against one another. Any discussion of how such harmful speech acts may potentially be sanctioned through international criminal law in these settings must, however, also be accompanied by a recognition of the vital counter role that freedom of speech can play – especially in conflict settings. Ensured access to free, public, and independent information and the protection of journalistic and academic expression remain crucial tools for combatting disinformation and hate speech.

21. As such, while recognising the harm which has been caused by disinformation and speech crimes on the one hand, it must also be clearly reiterated that the solution is not found in an overly broad restriction on the right to freedom of speech and expression. Where international criminal law is used as a framework to sanction prohibited speech, it remains crucial that international courts do not infringe upon the standards for protecting freedom of speech and expression established through international human rights law. In the case of *Nahimana, Barayagwiza and Ngeze*, the ICTR noted that international human rights law 'has been sufficiently well developed in the areas of freedom from discrimination and freedom of expression [as] to be the point of reference for its consideration of these issues, noting that domestic law varies widely while international law codifies evolving universal standards'.[2]

2 *The Prosecutor v Ferdinand Nahimana, Jean-Bosco Barayagwiza, and Hassan Ngeze* (Judgement and Sentence) ICTR-99-52-T (03 December 2003) § 1010 – hereafter also referred to as *Nahimana et al.* Also discussed in Wilson and Gillett (Key Guidelines, n 1) § 22.

1 Speech Act Theory

22. Before delving into the more specific discussions and topics outlined above, it is necessary to briefly define what constitutes a speech act and to consider how speech act theory can provide a useful conceptual tool for analysing speech crimes in the context of international criminal law. This is important because speech act theory provides a foundation for distinguishing between those aspects of a speaker's utterance which themselves constitute an action and the 'consequences that rely on a separate actor or other circumstances'.[3]

23. Speech is communicated when a speaker performs an utterance which is heard by an individual or audience.[4] Speech act theory contends that beyond the mere communication of information, speech can also function as different forms of verbal actions used by the speaker to accomplish something.[5] It provides a framework through which to analyse and understand a speaker's intention by considering not just the literal meaning of the utterances performed but also the function and action they convey directly or indirectly. For implicit messages or veiled speech, the influence of a speech act on its receiving audience may be heavily dependent on their interpretation and its understood meaning within the social, political, and cultural context within which it was uttered. Nonetheless, if the intended impact is contained in the message, whether implicitly or in a veiled form, the statement maker can potentially be held legally responsible for it (and for any harmful impacts it may have), subject to the demonstration of the required elements of a crime and/or mode of liability.[6] For direct and public incitement to genocide, the *Akayesu* Trial

3 Shannon Fyfe, 'Tracking Hate Speech Acts as Incitement to Genocide in International Criminal Law' (2017) 30:2 Leiden Journal of International Criminal Law 523, 533.

4 Or otherwise received. The term 'heard' here can be understood broadly to also include written speech which is read.

5 Speech act theory was initially developed by JL Austin, *How to do Things with Words* (HUP 1962, further edition 1975) among others.

6 In relation to messages that are not overt in their aims, the 'indirect' label that may be applied potentially clashes with the form of liability set out under Article 25(3)(e) of the Rome Statute (Key Guidelines, n 2), which prohibits direct and public incitement to genocide. However, here it should be clarified that the 'direct' element does not exclude implicit messaging, and each case must be assessed based on its own facts and context. In the *Akayesu* case at the ICTR, the Trial Chamber clarified that although 'more than mere vague or indirect suggestion' is required to fulfil this element 'the direct element of incitement should be viewed in the light of its cultural and linguistic content. Indeed, a particular speech may be perceived as "direct" in one country, and not so in another, depending on the audience. The Chamber further recalls that incitement may be direct, and nonetheless implicit. The Chamber will therefore consider on a case-by-case basis whether, in light of the culture of Rwanda and the specific circumstances of the instant case, acts of incitement can be viewed as direct or not, by focusing mainly on the issue of whether the persons for whom the message was intended

Chamber said that the essential question is whether the intended audience would grasp the implications of the message.[7]

24. Speech act theory – which differentiates between the locutionary, illocutionary, and perlocutionary dimensions of a speech act – provides an important starting point for seeking to better understand and address some of these uncertainties as it provides the conceptual tools for distinguishing different types of speech acts and understanding their functions.[8]

Speech act theory classifies speech acts into different categories which are analysed on different levels:

- *A locution – in brief, the meaning and content of the speech act*; this is the performance of an utterance/statement 'in the full normal sense'.[9] In other words, locutions are centred around the speaker's utterances and what they ordinarily mean to a community of speakers in that specific time and place (in short, the meaning).
- *An illocution – in brief, the encouragement or exhortation of the speech act*; this goes further than the act of saying something, instead referring to the intent of, or goal sought by, the speech act; in other words, the active result it seeks to evoke through its implied meaning or request. For example, the illocutionary dimension of the speech act, 'I warn you that dog is dangerous' is to encourage the listener not to approach the dog (in short, the encouragement or incitement).
- *A perlocution – in brief, the impact of the speech act*; refers to the effect which is produced by the locutionary or illocutionary act on the listener – regardless of what the intention of the speaker may be. Examples of perlocutionary acts include convincing, encouraging, scaring, inspiring etc. These depend on the felicity conditions of the speech act, or contextual conditions that influence the uptake of the illocution on the part of the audience of the speaker (in short, the consequence of the speech act).

Of course, when analysing a potential speech crime, the existence of intent to elicit a specific effect on the listener is relevant to questions of criminal accountability in several respects. This is referred to and discussed throughout in Chapter 1 and can be particularly complicated for cases where the meaning of speech may be veiled (i.e. indirect) or hybrid (combining elements of being both direct and indirect) in nature.

---

immediately grasped the implication thereof.' *Prosecutor v Akayesu* (Judgement) ICTR-96-4-T (02 September 1998) §§ 557–558.

7 *Akayesu* TC (Introduction, n 6) §§ 557–558.

8 As articulated in Austin (Introduction, n 5) 94–101; For an application of speech act theory to international speech crimes see also Richard A Wilson, 'Inciting Genocide with Words' (2015) 36 Michigan Journal of International Law 277.

9 Austin (Introduction, n 5) 94.

CHAPTER 1

Interpreting and Analysing Speech Acts

25. The adjudication of alleged speech crimes necessarily involves interpreting the speech acts and analysing their content within the applicable legal framework. There are several key challenges associated with interpreting speech, including the difficulties inherent to the translation process and the importance of culturally informed analysis of language. The approach taken in key international trials, including *Nahimana et al.* and *Akayesu* at the ICTR as well as the *Krstić*, *Mladić*, *Karadžić* and *Popović et al.* at the ICTY demonstrates the need for a contextual analysis of speech and culture-specific utterances.[1] This has implications for evaluating the directness of a message, which according to the ICTR case law considered below does not have to be explicit to qualify as direct and public incitement to genocide. Implicit messages or utterances may also amount to incitement if those who are addressed immediately grasp their direct implications (i.e. as encouragements, or illocutions) due to their sociolinguistic content and cultural meaning.[2]

26. There is a general consensus in the literature that culture-specific utterances were explained and contextualised effectively in ICTR-related trials.[3] The Tribunal's consideration of the meaning and effect of speech acts emphasised the importance of contextualised analysis of such terms and their use by instigators and direct perpetrators leading up to and during the Rwandan genocide. Successive ICTR Trial Chambers relied on reports and testimony by Rwandan linguistic experts[4] to comprehend the meaning of the Kinyarwanda terms, including proverbs and local expressions.[5]

27. For example, in *Nahimana et al.* (the 'Media Case'), linguistic experts provided crucial testimony about Kinyarwanda terms and their culturally informed analysis of language was central to the case and supported 'the

1 This is also reviewed in the Hartford Guidelines. See Wilson and Gillett (Key Guidelines, n 1) §§ 274–289.

2 Wilson and Gillett (Key Guidelines, n 1) § 282.

3 Wilson and Gillett (Key Guidelines, n 1) § 282.

4 See Wilson and Gillett (Key Guidelines, n 1) §§ 277–280 for further explanation regarding the importance of the testimony provided by linguistic expert Dr. Mathias Ruzinda of the University of Rwanda and the report and testimony of Rwandan linguist Dr. Evariste Ntakirutimana.

5 Wilson and Gillett (Key Guidelines, n 1) § 282.

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-003
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

inference that calls [on the RTLM radio] to exterminate the *inkotanyi*[6] and *inyenzi* (cockroach) referred to Tutsis in general as opposed to simply their military forces' and thus 'constituted a form of genocidal incitement to eradicate Tutsis as a group'.[7] The 'directness' of these specific utterances (i.e. their locutionary and illocutionary dimensions) was identified and analysed in the ICTR investigations and trial proceedings according to their 'implicit content and meaning, and the way they were directly understood by their intended audiences'.[8]

28. The question of directness and inferring genocidal intent from derogatory speech, alongside other evidence, also arose during key trials at the ICTY. For example, in the trial of General Ratko Mladić, the Trial Chamber considered the defendant's use of the historically charged pejorative term *poturica*[9] to refer to Bosnian Muslims at a time when his troops where committing mass crimes against them to imply discriminatory intent.[10]

29. When interpreting the meaning (or locutionary aspect) of terms used by defendants accused of instigating or perpetrating atrocity crimes, it is crucial to consider that meaning is dependent on time and place. But time and place in the digital era is no longer restricted to a physical time or place in the traditional sense. Instead, it can encompass all manner of virtual platforms, such as the metaverse, along with disparate geographical locations. Consequently, the relevant contextual interpretive matrix may include virtual environments in which both natural and virtual and even agentic AI entities exist. The range of languages, dialects, and frameworks of communication have expanded in

---

6 The extended meaning of this term is explained by linguistic expert Dr. Mathias Ruzinda in his testimony at the ICTR as 'including RPF sympathiser or supported, and in some instances, it even seemed to make a reference to Tutsi as an ethnic group'. See *Akayesu* TC (Introduction, n 6) §§ 147, 365. See also Wilson and Gillett (Key Guidelines, n 1) § 277.

7 See Wilson and Gillett (Key Guidelines, n 1) § 279 which explains the importance of Ruzinda's testimony and his interpretation of the Radio Télévision Libre des Mille Collines (RTLM) radio broadcasts in shaping this view.

8 Wilson and Gillett (Key Guidelines, n 1) § 282.

9 This meaning of the term *poturica* (referring to the converts to Islam but also connoting a renegade or traitor) and the importance of considering it against its historical use and context is discussed in Wilson and Gillett (Key Guidelines, n 1) §§ 286–289.

10 As noted in Wilson and Gillett (Key Guidelines, n 1) § 284, context is not the only factor which determines the assessment of derogatory speech acts and there may be significant differences in the approach taken by judges of international tribunals towards inferring genocidal or discriminatory intent. For example, in *Krstić* the Appeals Chamber agreed with the Trial Chamber's assessment that 'charged language is commonplace amongst military personnel during war' but was not satisfied with the conclusion that genocidal intent could be inferred, finding instead that 'no weight can be placed upon Radislav Krstić's use of derogatory language in establishing his genocidal intent.' *Prosecutor v Radislav Krstić* (Judgement) IT-98-33-A (19 April 2004) § 130.

the online environment. Even if English has become almost ubiquitous, there is an ever-growing range of ways in which English and other languages are communicated, whether through emojis, acronyms, slang, or a combination of these modes. Building on this, the current PJI Guidelines expand and update international criminal proceedings that took place since the publication of the Hartford Guidelines, along with providing additional depth on some cases that preceded the current digital era. Based on those further reviews, additional sub-topics on the theme of analysing the contents of major speech acts have emerged, as set out below.

30. While international courts have engaged in contextual analysis when interpreting the meaning of terms in cases where the charges include incitement to genocide and hate speech as persecution, we extend their reasoning here to analyse orders in the context in which they are given.[11] Our discussion also explains the notion of linguistic justice, which is vital in international settings where the language at trial may not be the language used during the conflict or societal setting in which the crimes occurred. The second section below considers how speech acts are assessed legally against the applicable normative standards, with multiple case studies presented to illustrate the analytical approach taken by judges at the international courts. Notwithstanding the overarching focus of these Guidelines on the digital era, this first chapter looks in detail at what could be considered more 'traditional' forms of speech crimes analysis, based on witness testimony concerning orders and courtroom interpretation. This provides foundational normative and jurisprudential guidance which informs the subsequent sections which address topics of particular relevance in the current digital age.

1 Linguistic Justice: the Interpretation of Speech in Context

31. The meaning of speech acts is seldom direct and obvious to all concerned and frequently requires additional interpretation by courts. As such, there are numerous challenges associated with interpreting facially ambiguous or veiled speech in speech crimes trials.[12] The fact that many international judges may 'neither speak the language of the region in question, nor have a thorough knowledge of its history and culture' risks misunderstandings concerning the possible range of meanings of the speech in question or fact witness

11 See Wilson and Gillett (Key Guidelines, n 1) §§ 71–73, 77, 138–139, 172, 255.

12 This is also acknowledged in the section on 'Expert Witnesses, cultural specificity and the meaning of words' in Wilson and Gillett (Key Guidelines, n 1) § 274.

accounts.[13] Consequently, there is the linguistic aspect of interpreting speech acts and the socio-cultural aspect of understanding their meaning which are inter-related.

32. On the technical linguistic side, there are terms that are so culturally and contextually bound that there may be no analogue in the Court's working languages. In this light, the ICTY translation service has noted that in some cases it may be preferable to 'leave terms in the original and annotate them' when no accurate and equivalent single interpretation or translation into English is feasible due to attached cultural or historical meaning.[14] In such instances, the parties will have to 'determine whether expert evidence is necessary to explain the specific context-based interpretation of specific words, in combination with other evidence of their impact and intent, including testimony from insider and crime-based witnesses'.[15]

1.1 Case Vignette: Ntaganda at the ICC

33. Additional reviews of the case law reveal that uncertainty about the meaning of a term may be exacerbated by contradictory witness accounts or where a party contests the explanations of a term provided by witnesses, including insider witnesses. For example, in the *Ntaganda* trial at the International Criminal Court, the parties contested the extended meaning of the defendant's order *kupiga na kuchaji* and his use of the term *kadogo* and they advanced contradictory interpretations of their use and meaning, as explained below.[16]

34. In *Ntaganda*, the defendant was indicted and convicted of 'conscripting or enlisting children under the age of fifteen' among other crimes.[17] The word *kadogo* was important in this respect as it was often used, including by the accused, and allegedly referred to child soldiers. The term came up in the accounts of multiple insider witnesses during both the investigation[18] and trial. Testimony about the presence of *kadogo* soldiers in the *Forces Patriotiques pour*

13 Wilson and Gillett (Key Guidelines, n 1) § 274.

14 Wilson and Gillett (Key Guidelines, n 1) § 285.

15 Wilson and Gillett (Key Guidelines, n 1) § 285.

16 Further discussion on this in Rosemary Grey and Ludmila Stern, '*Kadago* in the Courtroom: Language Disputes in Atrocity Trials' (*Opinio Juris* 31 January 2024) <https://opiniojuris.org/2024/01/31/kadagos-in-the-courtroom-language-disputes-in-atrocity-trials/\> accessed 26 June 2026.

17 Rome Statute (Key Guidelines, n 2) art 8 War Crimes 2(e)(vii) and as an indirect co-perpetrator under art 25(3)(a).

18 See *Prosecutor v Bosco Ntaganda* (Decision Pursuant to Article 61(7)(a) and (b) of the Rome Statute on the Charges of the Prosecutor Against Bosco Ntaganda) ICC-01/04-02/06 (09 June 2014) § 91, 95 – footnote 381.

la Libération du Congo (Patriotic Forces for the Liberation of Congo, or FPLC) and among Ntaganda's escorts played a key role in proving the recruitment and use of child soldiers, of which Ntaganda was ultimately convicted.

35. When asked to explain the meaning of the term *kadogo*, numerous insider witnesses consistently equated it with 'child soldiers' and stated that the term would not have been used to refer to someone over eighteen.[19] Meanwhile, Ntaganda himself offered a less incriminating meaning of the term – explaining that 'the use of the term *kadogo* within the *Union des Patriotes Congolais* [Union of Congolese Patriots] (UPC)/FPLC was not related to an individual's age, but rather referred to a person's size'.[20] However, in light of the numerous and consistent witness testimonies that *kadogo* 'was used to refer to 'children' or very young soldiers', the ICC Trial Chamber was unconvinced by Ntaganda's claims about the meaning of *kadogo*.[21]

36. Another linguistic issue addressed in the *Ntaganda* proceedings was the meaning of *kupiga na kuchaji*, an order given by Ntaganda and other commanders to FPLC troops, which is discussed in greater detail below in the section 'Orders as Speech Crimes'. The Prosecution alleged that this essentially was an order to attack and kill civilians and soldiers alike and to destroy and loot their property. The Defence in *Ntaganda* contested this interpretation of the phrase. According to Ntaganda, this meant only to 'charge at the fleeing enemy after having beaten them, and to take all their equipment' but which he explained 'did not involve looting, because it was forbidden to steal from the civilian population'.[22] Ntaganda also denied having himself ever used the expression.[23]

37. Ultimately, the Chamber did not find credible Ntaganda's proposed interpretation of this phrase given the other evidence it had received on this matter and instead found that the phrase did mean to target civilians and soldiers alike.[24] To support its conclusion on the meaning of the expression *kupiga na kuchaji*, the Trial Chamber evaluated evidence that included testimony by multiple witnesses who provided a 'clear', 'unambiguous' and 'consistent' interpretation of the phrase.[25] The Trial Chamber's finding on the meaning of the phrase *kupiga na kuchaji* was affirmed by the Appeals Chamber which highlighted that 'the order must be interpreted in context' and that in relation

19 See *Prosecutor v Ntaganda* (Judgement) ICC-01/04-02/06 (08 July 2019) § 359 and accompanying footnote 995.

20 *Ntaganda* TC (Chapter 1, n 19) § 359, referring especially to accompanying footnote 994.

21 *Ntaganda* TC (Chapter 1, n 19) § 359, footnote 994.

22 *Ntaganda* TC (Chapter 1, n 19) § 415.

23 *Ntaganda* TC (Chapter 1, n 19) § 415.

24 *Ntaganda* TC (Chapter 1, n 19) § 415.

25 *Ntaganda* TC (Chapter 1, n 19) § 415. See also accompanying footnotes 1186 and 1188.

to this order 'crimes against civilians were subsequently committed'.[26] The Appeals Chamber therefore rejected Ntaganda's appeal on the exceptionality and contested meaning of the phrase.

38. The Trial Chamber interpreted the specific words used by the defendant in their context of their usage, including the relevant cultural context. It also considered the military context and more broadly the events that were unfolding at the time. The Trial Chamber did not receive any expert evidence specifically focused on the meaning of *kupiga na kuchaji*. But contextual evidence about the background to the conflict was provided by prosecution witness Roberto Garretón, a legal expert who served between 1994 and 2001 as UN Special Rapporteur on the Situation of Human Rights in Zaire/DRC. Given the contested meaning of this phrase, and its centrality to the litigation, the *Ntaganda* case provides an example of the importance of gathering evidence from witnesses who operated closely with the accused regarding its meaning. Because the language in question has a relatively small population of speakers, there will be few linguistic experts who can testify about the meaning of specific phrases. In these circumstances, courts may ground their interpretation in alternative and diverse sources of evidence.[27]

39. Looking at the methodology for assessing speech acts during international litigation, as set out in the Key Guidelines herein, a distinction can be made between the linguistic analysis of the meaning of a speech act and the judicial analysis of the significance of the speech act in relation to proving a crime or an element of a crime. Regarding linguistic analysis and the interpretation of speech, it is worth elaborating further on best practices when there are contesting accounts on the meaning of legally significant culture or context specific utterances in languages of lesser diffusion. This is especially acute when utterances are considered to have the potential to have contributed to the commission of core international crimes and as such, their accurate interpretation plays a key role at trial. This raises the question of whether and how the rarity of a language and its diffusion (if, for example, a language is spoken

26 *Prosecutor v Ntaganda* (Public redacted version of Judgement on the appeals of Mr Bosco Ntaganda and the Prosecutor against the decision of Trial Chamber VI of 8 July 2019 entitled 'Judgement') ICC-01/04-02/06 A A2 AC (30 March 2021) § 505.

27 Also referred to in linguistic circles as 'languages of lesser diffusion' – meaning that they may 'have few speakers in a given location, are non-standardised and largely oral, and for which few or no accredited interpreters exist'. For further discussion on the logistical challenges of interpreting these languages in an international court setting refer to Leigh Swigart, 'Unseen and Unsung: ICC Language Services and their Impact on Institutional Legitimacy' in Freya Baetens (ed) *Legitimacy of Unseen Actors in International Adjudication*, edited by Freya Baetens (CUP 2019) 272.

in only a certain more isolated location or by a limited number of speakers or exists only in the oral tradition) impacts the contribution of an utterance in said language to the commission of an atrocity crime.

40. For the digital era, these lessons have profound ramifications. The context in which speech acts must be interpreted include social media platforms, where messages may be conveyed in a wide range of ways, from typed words, to emojis, to video and audio messages, to likes and other forms of engagement. The distinction between the linguistic analysis and the judicial analysis is all the more important in such contexts. If the two are conflated, then the legal assessment can influence the interpretation of the speech as a factual matter. Equally, factual considerations may be merged with legal considerations, muddying the clarity and certainty of the law. Ultimately, this could potentially lead to improper findings by trial chambers.

1.2 Case Vignette: Krstić and Popović et al. at the ICTY

41. An important example of linguistic analysis from the ICTY cases on the former Yugoslavia, is the term *asanacija*. This term can refer to cleaning, cleansing, or other variations of these concepts. It was used frequently in orders and other communications, meaning that its meaning was particularly relevant in determining criminal liability. The disputed meaning of this term (including as it related to the genocide at Srebrenica) is discussed in both the *Krstić* and *Popović et al.* Judgements and the ICTY translation expert Ellen Elias-Bursać referred to *asanacija* in the following terms:[28]

> While several witnesses use a language query to bolster the credence of their testimony and the defence uses arguments over translation and interpreting to query responsibility and raise reasonable doubt, the prosecution has its own approach to the 'struggle over meaning'. They used the disputes over the word *asanacija* and its translations to provoke and unsettle defence witnesses. The judgements assert that *asanacija* was a code word used by the Bosnian Serbian military units to signify the activity of burying bodies to hide a massacre.[29]

42. By its nature as international law, international criminal justice is marked by significant differences in language, worldview, and cultural practices between

28 Ellen Elias-Bursać, *Translating and Interpreting Evidence at a War Crimes Tribunal: Working in a Tug of War*, (Palgrave Macmillan 2015) Chapter 5 'The Office of the Prosecutor' 181–193.

29 Elias-Bursać (Chapter 1, n 28) 181–193.

those administering proceedings and those appearing before the court.[30] This can lead to a number of challenges, especially when speakers of less well-known languages require accommodations so that there is due process, raising particular concerns about the accuracy of interpretation and the fairness of a trial.[31] This constellation of translation-related issues has been referred to under the concept of 'linguistic justice'.[32] Such difficulties may arise at any point in an international trial, including in the gathering of evidence but also throughout the trial proceedings and onto the appeal and any possible subsequent review – as discussed in greater detail below.

43. While such language issues are relevant for all crimes investigated and prosecuted at the international level, they are of particular relevance to the topic of speech crimes. This section discusses potential linguistic justice challenges when speech crimes play a role in an international trial, including potential accountability gaps and possible fair trial concerns.

1.3 Case Vignette: Interpretation at the Special Court for Sierra Leone

44. Several such fair trial concerns related to the challenges of multiple layers of translation emerged at the Special Court for Sierra Leone (SCSL). Although the Court's official language for proceedings was English, Sierra Leone is a multi-lingual country in which its citizens speak more than twenty-four languages. The Court offered translation into local languages such as Krio and Mende among others.[33] Despite efforts by the Court to provide multi-lingual interpretation and make its proceedings accessible through different formats, pervasive challenges remained which are important both for the rights of victims and the accused. These challenges impacted not only the proceedings themselves but also the public understanding and perception of proceedings.

---

30 Leigh Swigart, 'Linguistic and Cultural Diversity in International Criminal Justice: Toward Bridging the Divide' (2016) 48 University of the Pacific Law Review 197.

31 Swigart (Chapter 1, n 30).

32 On the general notion of 'linguistic justice' see Philippe Van Parijs, *Linguistic Justice for Europe and for the World*, (Oxford University Press 2011) Chapter 4 'Linguistic Justice as Parity of Esteem' 117–132. In the context of the ICTY, see Besmir Fidahić, *Linguistic Justice at the International Criminal Tribunal for the Former Yugoslavia* (Cambridge Scholars 2020).

33 Rebekka Ehret, 'Translation services at international tribunals – What happens if social pressures conflict with professional ethics?' (2009) TRIAL: Die Zeitschrift der Schweizerischen Gesellschaft für Völkerstrafrecht 2009; Rebekka Ehret, 'Interpreting from and to Non-Standardised Languages at the Special Court for Sierra Leone' (*1 Trial 4 Languages*, 2004) <https://1trial-4languages.org/sierra-leone/interpreting-from-and-to-non-standardised-languages-at-the-special-court-for-sierra-leone/\> accessed 24 June 2026.

45. Critical observers such as Human Rights Watch contended that there was a lack of experienced and trained interpreters in the numerous local languages – some of which existed primarily as oral traditions – meaning that field translators, investigators, and Court interpreters were often relied upon based on availability. Moreover, there was a lack of standardised practice in the early phases of investigations and proceedings.[34] The fact that the legal terminology employed by the Court in English did not always have an equivalent in the local languages further compounded concerns around the quality and lack of standardisation in translation and interpretation.[35] In some cases, the semantic meaning of a word in the local language could be reasonably expanded to approximate the standard English legal term. In other cases, the English word was taken as a loan word or if it was not likely to be understood, 'the legal term was simply explained'.[36]

46. In the experience of the SCSL, the issues highlighted above are not specific to speech crimes per se but emerged in the context of general challenges relating to translation and interpretation at international criminal tribunals. However, such concerns may be amplified in cases where a speech crime element is present. Given the central importance of understanding and placing the meaning (or locution) of an utterance within its social, political and historical context, there is a need to ensure nuanced meaning (i.e. when a term may have legal significance, particularly where this may be veiled and not immediately evident at face value) is not lost when there are multiple layers of translation.

47. More broadly in relation to public outreach, although the SCSL made various efforts to render its proceedings public and even provided regular radio interpretations thereof, a significant portion of the Sierra Leonian population were prevented from engaging with the Court due to language barriers.[37] Indeed, studies conducted to examine the effectiveness of the Court's outreach efforts found that although general awareness of the Court itself was high, the depth of understanding of the Court's precise role was low.[38] Without

34 Human Rights Watch (HRW), 'Bringing Justice: The Special Court for Sierra Leone – Accomplishments, Shortcomings, and Needed Support' (Human Rights Watch, 8 September 2004) <https://www.hrw.org/report/2004/09/08/bringing-justice-special-court-sierra-leone/accomplishments-shortcomings-and\> accessed 16 June 2026.

35 Ehret, 1 Trial 4 Languages (Chapter 1, n 33).

36 Ehret, 1 Trial 4 Languages (Chapter 1, n 33).

37 Ehret TRIAL (Chapter 1, n 33).

38 Rachel Kerr and Jessica Lincoln, 'The Special Court for Sierra Leone: Outreach, Legacy and Impact' (*War Crimes Research Group, King's College London* 2008) <https://www.rscsl.org/Documents/slfinalreport.pdf\> accessed 16 June 2026.

recognising the hybrid nature of the Court, and as such the role of the government and society of Sierra Leone in its functioning, there was a perception among many participants that the law used at the Court represented an imposition of 'white man's law'.[39]

1.4 Case Vignette: the Importance of Socio-Linguistic Experts for Fair Trials – a Lesson from Western Australia

48. An example from the Australian domestic legal system which provides cautionary guidance regarding linguistic justice for international or hybrid tribunals, is the 2003 case *Murray v Feast* from the state of Western Australia. It highlights the importance of having not only interpreters skilled in language and translation but also socio-linguistic experts. There are lessons which can be drawn from this case of relevance also to speech crimes in international criminal law, as it illustrates the risks of potential fair trial violations that may occur in the absence of a socio-linguistic and contextual approach.

49. At the expense of ancestral languages like Kriol or Aboriginal English, the Australian legal system, including judicial proceedings, operates in Standard Australian English.[40] While it has been recognised by the Australian Courts that criminal trials concerning a defendant who does not speak the language in which the proceedings are being conducted and where an interpreter is not provided will result in an unfair trial,[41] the case of *Murray v Feast* illustrates a unique issue that emerges when two different languages use what appear to be identical terms or phrases but employ them in a way that accords them different meanings. At trial, the defendant, Murray, was convicted of one charge of aggravated common assault. It was alleged that he kicked a 70-year-old man in the genitals in October 2020. Murray, an aboriginal man, gave evidence at trial with the assistance of an interpreter who was instructed by the Magistrate to not translate portions of evidence which appeared to be in Standard Australian English.[42] Although informed by both the interpreter and Murray's counsel that Murray was speaking in Aboriginal English and Kriol, which therefore necessitated interpretation due to the possible distinct meaning of certain words and phrases even when they might appear identical in English, this was refused. When the Magistrate presiding over the case heard Murray using English words, he dismissed the interpreter. Murray's lawyer sought to

39 Kerr and Lincoln (Chapter 1, n 38).

40 Aidan Ricciardo, 'Denial of Language: Grounds for Occasioning an Unfair Trial?' (2024) 1:3 First Nations Law Bulletin 43, 43.

41 *Ebatarinja v Deland* [1998] HCA 62 [26]-[27].

42 *Murray v Feast* [2023] WASC 273 [25].

have the ruling overturned, arguing: 'my understanding is that a word in Kriol, whilst it might sound English, could actually have a different …', before being cut off by the Magistrate. However, the Magistrate stuck with his ruling and Murray was convicted and given a 10-month suspended sentence.

50. On appeal, the first instance trial of Murray was ultimately found to be unfair by Judge Solomon of the Supreme Court of Western Australia on the basis that his 'linguistic distinctiveness' had not been recognised by the Magistrate.[43]

51. The refusal to allow interpretation of the testimony was found to be indicative of 'the hallmarks of at least the impression of an unfair prejudgment of Mr Murray's character and the integrity of his testimony' thereby resulting in an unfair trial.[44] Furthermore, in assuming that the words spoken by Murray in Aboriginal English or Kriol had the same meaning in Standard Australian English the Magistrate had denied the linguistic distinctiveness of the appellant, including by failing to take into account all of the relevant historical and social context and circumstances which had come to shape this distinction.[45] The decision of the Magistrate and the apparent biased attitude it stemmed from was thus not compatible or consistent with the appearance of impartiality.

52. The Judgement of the Supreme Court of Western Australia in *Murray v Feast* also raises germane questions about the admissibility of expert reports on sociolinguistics. Here the Judgement states that whereas 'an ultimate conclusion about the credibility or truthfulness of a particular witness is a matter for the factfinder and is, therefore, not a proper subject for expert opinion' […] 'an expert may give evidence on the language and communication difficulties which are likely to be experienced by an Aboriginal witness'.[46] It was further noted that 'expert comments on particular parts or passages of evidence given by witness at trial are likely to usurp the role of the fact finder – that is, those comments typically relate directly to the evaluation of testimony, which is an ultimate issue for the Court'.[47]

53. Based on the foregoing, in certain contexts there may be a need not just for language or linguistic experts but also for sociolinguistic experts to

43 Ricciardo (Chapter 1, n 40).

44 *Murray v Feast* (Chapter 1, n 42) [175].

45 See Aidan Ricciardo, 'An unfair trial occasioned by denial of linguistic distinctiveness and an appearance of pre-Judgement: Murray v Feast [2023] WASC 273' (*Evidence Law in Western Australia* 2023) <https://waevidence.com/2023/08/18/an-unfair-trial-occasioned-by-denial-of-linguistic-distinctiveness-and-an-appearance-of-preJudgement-murray-v-feast-2023-wasc-273/>.

46 Ricciardo (Chapter 1, n 45).

47 Ricciardo (Chapter 1, n 45).

guarantee a fair trial – although the Court will always retain the authority to evaluate evidence and testimony. Prior social research finds that international criminal tribunals have been receptive to expert witness testimony by anthropologists, historians, and linguists, about the language and history of a country or region in which the tribunal is adjudicating crimes.[48] Additionally, the paradigm raises the question of the independence of linguistic experts. Given that judicial independence is required at the ICC (as per Article 40 of the Rome Statute) and is elaborated in its jurisprudence, it raises the question of whether bespoke standards should be set out to ensure the independence of interpreters.

54. For speech crimes, it is important that witnesses' words are accurately interpreted, particularly when the accused is testifying. That interpretation should be culturally sensitive, as some languages may use English terms but involve meanings that differ from the standard English meanings. Moreover, whereas experts may provide evidence on language and the communication difficulties experienced by the accused, they should not opine on the overall credibility of other witnesses, including the accused, as that would encroach on the judges' authority to determine such issues.

55. Although the questions highlighted above are not specific to speech crimes per se, but relate to general challenges around translation and interpretation, they may become more acute in cases which involve a speech crime element. Given the need to understand the meaning of an utterance in its social, political, and perhaps even historical context, there is a need to ensure that nuanced meaning (i.e. when a term may have legal significance, particularly where this may be veiled and not immediately evident at face value) is not lost when there are multiple layers of translation. Moreover, in the context of the digital era, there are sub-cultures that create and use languages that may be based on existing languages but are not comprehensible to the general public who do not interact with these groups. In this light, the availability of linguistic experts for cases hinging on speech acts should be expanded to include experts with knowledge of the specific digital sub-culture or linguistic currency of the specific group under the spotlight in criminal proceedings.

2 Forensic and Legal Analysis of Speech Acts

56. Having established the semantic meaning of the statement in question (whether it is in writing or verbal), our analysis turns to the forensic

48 Richard A Wilson, 'Expert Evidence on Trial: Social Researchers in the International Criminal Courtroom' (2016) 43:4 American Ethnologist 730.

significance of an expression and whether it can be a basis for a conviction as a speech crime (either a substantive crime such as persecution, or a mode of liability). The following section examines how speech can be analysed through the normative framework applicable to cases before the ICC, with a particular focus on orders which have been found to constitute speech crimes.

2.1 Orders as Speech Crimes

57. The examples before the ICTR and ICTY noted above raise additional questions about how to interpret the implications of other kinds of context-specific speech acts that encourage or direct others to commit atrocity crimes (or illocutions). This section will focus on interpreting and understanding the implicit contextual meaning of orders.

58. The unique approach of framing of ordering as a mode of liability in international criminal law is based on three elements which must be fulfilled.[49] Ordering as a mode of criminal liability therefore requires (i) a position of authority; (ii) The *actus reus* (the physical element) consisting of instructing or commanding a person to perform an act;[50] (iii) an effect on the crime in question; in addition to the *mens rea* (the mental element). Further review of recent key cases, in particular the *Ntaganda* Trial Judgement at the ICC and the *Mladić* and *Karadžić* Appeals Judgements at the International Residual Mechanism for Criminal Tribunals (IRMCT), highlights another dimension to consider when evaluating and interpreting orders – the relevance of context for analysing the meaning of orders.

59. When assessing and contextualising orders, it must be considered whether they have made the likelihood of an atrocity crime being committed 'significantly' more likely either by intentionally, directly, and/or publicly inciting others, who are subordinate to them, to commit the crime. In other words, the order must have had a 'direct and substantial effect' on the commission of the crime(s). There are two elements relevant to assessing the language and contribution of orders. First, there needs to be an evaluation of the directness dimension of the order and whether the meaning of the order is clearly understandable for the recipient. Second, a court must consider where the speech

49 Wilson and Gillett (Key Guidelines, n 1) § 146, Section 5.2.1 Ordering.

50 The *actus reus* pertaining to ordering is discussed in Wilson and Gillett (Key Guidelines, n 1) §§ 152–157. It consists of instructing or commanding a person in a subordinate position to perform an act. This is set out in *Prosecutor v Dario Kordić and Mario Čerkez* (Judgement) IT-95-14/2-A (17 December 2004) § 28 and referred to in *Prosecutor v Jean-Pierre Bemba Gombo, Aimé Kilolo Musamba, Jean-Jacques Mangenda Kabongo, Fidèle Babala Wandu, and Narcisse Arido* (Judgement pursuant to Article 74 of the Statute) ICC-01-/05-01/13 (19 October 2016) § 77, hereafter also referred to as *Bemba et al.*

act lies on the spectrum of implicit to explicit – noting that explicitness is not a required element for any speech crime, while directness is.

2.2 Case Vignette: Ntaganda ICC Trial Judgement at the ICC

60. In its jurisprudence, the ICC elaborated the elements of the mode of liability of ordering contained in Article 25(3)(b) of the Rome Statute. For example, in *Al-Rahman*, which is summarised below, the Trial Chamber explained that '[a]s regards the mode of liability of "ordering", the *actus reus* means that a person in a position of authority instructs another person to commit an offence'.[51]

61. Guidance can also be gleaned from a Judgement which did not specifically involve ordering as a mode of liability,[52] but focused factually on orders that were issued by the Accused.[53] This is the case of *Ntaganda*, which includes a detailed discussion on the meaning of language contained within specific orders and whether or how this influenced the likelihood of an order contributing to the commission of an atrocity crime.

62. Bosco Ntaganda, a commander of operations of the FPLC, was indicted by the ICC Prosecutor for crimes against humanity and war crimes. The common and repeated use of the expression *kupiga na kuchaji* by Ntaganda and other commanders of the FPLC to give orders to attack played a key role in the Chamber's conviction of Ntaganda for crimes including intentionally attacking civilians, rape, pillage, ordering the displacement of the civilian population, and destroying the adversary's property.

---

51 *Prosecutor v Ali Muhammad Ali Abd-Al-Rahman ('Ali Kushayb')* (Judgement) ICC-02/05-01/20-1240 (06 October 2025) §§ 802–806. See also *Prosecutor v Bosco Ntaganda* (Decision on the Confirmation of Charges) ICC-01/04-02/06 (09 June 2014) § 145.

52 Ntaganda was convicted at trial as an indirect-co-perpetrator and as a direct perpetrator, that is, under Article 25(3)(a). See *Ntaganda* AC (Chapter 1, n 26) § 29. The issue of the orders given by the accused in this context related to the crimes he was charged with, namely, Rome Statute (Key Guidelines, n 2) art 8(2)(e)(viii) ('ordering the displacement of the civilian population for reasons related to the conflict, unless the security of the civilians involved or imperative military reasons so demand'); Rome Statute (Key Guidelines, n 2) art 8(2(e)(i) ('intentionally directing attacks against the civilian population as such or against individuals not taking direct part in hostilities'); and Rome Statute (Key Guidelines, n 2) art 8(2)(iv) ('intentionally directing attacks against buildings dedicated to religion, education, art, science or charitable purposes, historic monuments, hospitals and places where the sick and wounded are collected, provided they are not military objectives'). As such, the relevant 'orders' were part of the substantive offences which Ntaganda committed under Rome Statute (Key Guidelines, n 2) art 25(3)(a), as well as his role in the commission of those crimes.

53 Johannes Block, 'Ordering as an Alternative to Indirect Co-Perpetration: Observations on the *Ntaganda* Case' (2022) 20:3 Journal of International Criminal Justice, 717.

63. The literal meaning of *kupiga na kuchaji* was contested at trial. To arrive at a conclusion about how to interpret the phrase, it was thus necessary to place and analyse it within the specific military context in which it was uttered and through the testimony of multiple insider witnesses regarding its meaning. Notably, although the phrase *kupiga na kuchaji* contains no explicit or direct reference to attacking civilians, it was understood to nevertheless implicitly call for such acts.

64. The order *kupiga na kuchaji* was explained by insider witnesses to mean 'to attack and to loot, […] taking possession of the opponent's belongings, but also any kind of property, including that belonging to "civilians,"'.[54] In addition to the explanations provided through witness testimony in *Ntaganda*, the Chamber also took into account the events which unfolded following the use of *kupiga na kuchaji*. It was observed that 'during the assaults that followed these orders [*kupiga na kuchaji*], civilians were murdered and raped. The senior commanders further ordered forcible transfer of the civilian population'.[55] It was also found that when orders were given, 'troops behaved as instructed; indeed, they committed different types of violent acts targeting the civilians specifically'.[56] To understand the meaning of the order thus required contextualising it within the time and space in which it was made.

65. The unfolding of military operations and events on the ground following the use of *kupiga na kuchaji* aligned with the explanation of its meaning provided by insider witnesses. This also lends further credence to the notion that directness can be derived from the implicit content of speech and the way in which intended audiences directly understand its meaning.[57] Even without an explicit order calling for attacks on civilian populations, the direct nature of *kupiga na kuchaji* can therefore be established from the reality that 'pursuant to this order from their commander, these UPC/FPLC soldiers committed crimes against the Lendu'.[58]

66. The Chamber in *Ntaganda* states that 'in giving orders for troops to attack using the term "*kupiga na kuchaji*", Mr Ntaganda necessarily meant for his subordinates to attack everyone and everything, targeting Lendu civilians and their property specifically. […] For the Chamber, all this demonstrates that Mr Ntaganda willingly sent his troops to attack and kill Lendu civilians,

54 *Ntaganda* TC (Chapter 1, n 19) § 415.
55 *Ntaganda* TC (Chapter 1, n 19) § 671.
56 *Ntaganda* TC (Chapter 1, n 19) § 688.
57 Previously discussed in Wilson and Gillett (Key Guidelines, n 1) in the context of incitement to genocide in Rwanda, see §§ 71–77.
58 *Ntaganda* TC (Chapter 1, n 19) § 841.

to engage in sexual violence against this population, and to loot and destroy their belongings'.[59] The Trial Chamber thus found that although there was no specific reference to civilians in the order it was 'understood by the soldiers to mean attacking all the Lendu, including civilians, and to loot their property' within the ongoing military context.[60]

67. Explicitness is not a required element for any speech crime. Consequently, the directness of an order can also be derived from implicit statements when their meaning is clearly understood by the recipient and results in a person of authority instructing or convincing another to commit a crime.[61] On this subject, the Appeals Chamber in *Ntaganda* considered defence counsel's argument that it was unnecessary to specify 'Lendu combatants'[62] when issuing orders 'to attack the Lendu' as this was self-evident according to the defence.[63] Ntaganda submitted that his responsibility 'cannot hang on a failure to add the word "combatants" to "Lendu" in a wartime context'.[64]

68. The Appeals Chamber found no merit in this submission. Instead, it noted that the words conveyed in Ntaganda's order 'must be interpreted on the basis of the entire factual findings and evidence' including that 'the UPC leadership designated the Lendu ethnic group, including civilians, as the enemy, songs were taught to recruits as part of their training, inciting soldiers to attack or kill the Lendu and that recruits would get everything, including women, and the UPC/FPLC soldiers committed crimes against Lendu civilians.'[65]

69. The Appeals Chamber upheld the finding of the Trial Chamber that 'Ntaganda in his instructions to the UPC/FPLC troops, ordered them to attack the Lendu without distinction'.[66] As such, the failure to specifically include the word 'combatant' in his order, when viewed in the proper context, was not an omission which denoted a lack of positive action. The order without the word 'combatant' was indeed positive action and its meaning was well understood by the recipients. As noted by the Prosecutor in the *Ntaganda* Appeals, 'the meaning of the words must be interpreted in the context of the evidence in the case "where ethnicity was found to be a relevant dimension to the conflict"'.[67]

---

59 *Ntaganda* TC (Chapter 1, n 19) § 1186–1187.
60 *Ntaganda* TC (Chapter 1, n 19) §§ 415, 688.
61 For further discussion on the fundamental elements of the crime of ordering see Wilson and Gillett (Key Guidelines, n 1) §§ 147–149.
62 *Ntaganda* AC (Chapter 1, n 26) §§ 991, 998.
63 *Ntaganda* AC (Chapter 1, n 26) § 994.
64 *Ntaganda* AC (Chapter 1, n 26) § 990.
65 *Ntaganda* AC (Chapter 1, n 26) § 998.
66 *Ntaganda* AC (Chapter 1, n 26) § 994.
67 *Ntaganda* AC (Chapter 1, n 26) § 991.

70. This example from Ntaganda, although made in relation to crimes with an element of ordering and modes of liability other than ordering, demonstrates the crucial importance of contextualising speech in order to derive its meaning rather than considering only the literal words uttered.

71. Additional guidance on the *actus reus* of ordering includes the fact that, in a military context, discipline and a climate of obedience usually means that an order will be complied with and the directness element fulfilled, whether or not the order is unlawful.[68] Moreover, 'instructing' requires a positive action by the individual in the position of authority: ordering cannot be committed by omission (so failing to order is not an applicable mode of liability).[69]

2.3 Case Vignette: Abd-al-Rahman ICC Trial Judgement at the ICC

72. In the *Abd-al-Rahman* case, the ICC similarly considered the defendant's accessorial liability for ordering the Janjaweed forces within the meaning of Article 25(3)(b) of the Rome Statute for the war crime of attack against a civilian population, the crime against humanity and war crime of murder and attempted murder, the war crime of pillaging, the war crime of destroying the enemy's property, the crime against humanity of other inhumane acts, the war crime of cruel treatment, the crime against humanity and war crime of rape, the war crime of forcible transfer of population, and the crime against humanity of persecution committed during the Janjaweed operation in Kodoom, Bindisi and surrounding areas in Darfur, Sudan between 15–16 August 2003.[70] ICC Trial Chamber I found Ali Muhammad Ali Abd-Al-Rahman, a leader of the Janjaweed forces, guilty of the above and under ordering as a mode of liability, discussed extensively his use of the term *amsah aksah* and its implied meaning.

73. Based on the reports of multiple witnesses, the term *amsah aksah* was well-known and frequently employed by both senior Government of Sudan (GoS) officials and field commanders as well as by Janjaweed leaders during the armed conflict in Sudan.[71] For example, the Chamber referred to Minister Harun's use of the phrase during a public meeting attended by Abd-Al-Rahman and other members of the Janjaweed. The Chamber found that the use of this term and 'the fact that it was being uttered by a GoS minister, effectively communicated to those present at the public meeting that *amsah aksah* and "wipe

68 Wilson and Gillett (Key Guidelines, n 1) § 156.
69 Wilson and Gillett (Key Guidelines, n 1) § 152.
70 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 999.
71 *Abd-Al-Rahman* TC (Chapter 1, n 51) §§ 363, 364.

out and sweep away" authorised *inter alia* rape, killing civilians and pillaging'.[72] As such, when the same phrase was used later by Abd-Al-Rahman to direct the Janjaweed prior to attacks, 'he and the Janjaweed were well aware of the meaning and significance of the phrases he used to instruct'.[73] This supports the finding that the crimes which followed were sanctioned by the State, and that the Janjaweed understood the implied meaning of *amsah aksah* as an instruction or licence to commit the crimes which ensued (i.e. the consequences of the speech act, or perlocutions).[74]

74. As such, in noting that the term *amsah aksah* was a well-known expression used in the context of armed conflict in Sudan, the Trial Chamber emphasised the value of analysing 'the precise and accurate meaning this expression had when it was used during the time of the charges, and in particular in the context of the charged operations'.[75] This is notable because it demonstrates the layered and complex nature of analysing speech for the purpose of international criminal law. While the broader cultural, historical, political, and military context all play a role in understanding both the explicit and implicit meaning of terms and phrases, this sociocultural analysis must then be part of the review of the specific charges in the indictment. As such, the precise circumstances at hand and events unfolding on the ground also factor into the analysis of the appropriateness of speech crimes charges as a legal matter.

75. The term *amsah aksah* was used by Abd-Al-Rahman and others 'during speeches when addressing their fighters and when giving orders to attack, to "encourage" them "into action" and "boost [their] morale"'.[76] Metaphorically, the term can be understood as meaning to 'wipe out', to 'clean', 'to sweep away', and to 'not leave behind any trace' – the latter of which also included by burning.[77] According to the evidence and testimony provided by witnesses, this could be inferred to mean 'killing', 'with no limit' or 'relentlessly' – in other words, *amsah aksah* was used to encourage and justify killing and horrific aggression without 'restrictions or limits that govern military and combat operations'.[78] Based on the evidence provided, the Chamber was satisfied that the 'the expression "*amsah aksah*", in its common use and by a common and

72 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 1005.
73 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 1005.
74 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 1005.
75 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 364, footnote 874.
76 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 364.
77 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 365.
78 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 365.

tacit understanding shared by all on the battlefield (Janjaweed and GoS Forces members alike), meant that fighters shall kill and destroy without any limits'.[79]

76. On one occasion in Arawala, Abd-Al-Rahman 'told his men *amsah aksah* and "[a]nything you see, you have to kill or wipe. So no man has to stay alive" and on another in Dembow Kabdy he said "[…] don't be cowards, move forward. Don't be afraid. *Amsah aksah* and never bring someone alive."'[80] Throughout the operations in Kodoom and Bindisi, when instructing the Janjaweed to carry out their attacks, he shouted such chants as '[w]ipe out, sweep away, *amsah aksah*' while at the same time being heard saying '[…] don't leave anyone behind, and bring no one alive'.[81]

77. Therefore, although the meaning of *amsah aksah* was to some degree metaphorical and tacit, the Trial Chamber determined that 'the content of those speeches, including the use of *amsah aksah*, constituted explicit orders that declared no quarter and encouraged Abd-Al-Rahman's subordinates to: 'take no prisoners and kill anyone, including civilians […]; destroy everything and 'wipe out' entire areas'; act without any limits; [engage in] looting and pillaging.[82] The term *amsah aksah* was frequently coupled with language used by GoS officials and Janjaweed leaders, including Abd-Al-Rahman, which 'dehumanised the non-Arab tribes targeted by the crimes' where they were described as objects and 'something that could be eaten', 'something that needed to be cleaned' or 'something that the fighters could do what they wanted with, with no limits'.[83]

78. The Trial Chamber also analysed the context of ongoing intergroup mass atrocities in which such speeches were made. Regarding the military context and events unfolding, the speeches sometimes coincided with 'distributions of equipment, weapons or money' or in other cases even 'immediately preceded attacks and lootings'.[84] They also played on the context of historical tribal tensions which existed between non-Arab and Arab tribes.[85] Indeed, those speeches, addressed to GoS and Janjaweed fighters who were specifically selected on the basis of their belonging to Arab tribes, 'in this specific context of pre-existing tribal tensions, and which dehumanised members of non-Arab tribes, encouraged, as intended, the commission of crimes against them'.[86]

---

79 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 366, see also footnote 884 which refers to the Defence's Final Brief § 272.

80 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 361.

81 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 361.

82 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 826.

83 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 828.

84 *Abd-Al-Rahman* TC (Chapter 1, n 51) §§ 362, 829.

85 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 829.

86 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 829.

79. Finally, the Trial Chamber considered Abd-Al-Rahman's role and position of superior authority. The Chamber determined that Abd-Al-Rahman had *de facto* authority over the Janjaweed and on some occasions over Government of Sudan Forces. He commanded these forces in operations and gave instructions to the Janjaweed to commence its attacks on the villages and townships by shouting chants like *Allahu Akbar*, *amsah aksah* and 'wipe out and sweep away'.[87] It took into account that he was together with Minister Harun, who used similar phrases including *amsah aksah*, which in effect authorised killing civilians and pillaging, particularly given the fact these terms were 'being uttered by a GoS minister'.[88]

80. The Trial Chamber's treatment of Abd-Al-Rahman's speech acts provides valuable guidance for future speech crimes litigation. It shows that the meaning of the term *amsah aksah*, although metaphorical in expression and implicit, conveyed explicit orders to commit murder, property destruction, looting, and pillaging. The context in which these terms were used, and the subsequent actions undertaken by those subordinates and other members of Abd-Al-Rahman's forces who heard them, were critical evidence that allowed the Trial Chamber to reach conclusive findings as to the meaning and impact of this term.

2.4 Case Vignette: Karadžić IRMCT Appeals Judgement at the ICTY

81. The principle of evaluating the intent behind orders not only through the actual words uttered but also within the context in which they are issued was upheld in the *Karadžić* Appeals Judgement. The Appeals Chamber indicated that the evaluation of an order may require taking into account different aspects of the broader context. For example, in *Ntaganda* the ICC Chambers evaluated and interpreted orders in the context of the events unfolding on the ground while in *Karadžić*, the ICTY Chambers considered the relevant political context and relations with the international community as germane to the practical effect of orders on shaping the reality on the ground.

82. Whereas the discussion about *kupiga na kuchaji* from Ntaganda is pertinent as an example of orders in the context of the commission of atrocity crimes, the analysis of orders in the *Karadžić* and *Mladić* cases is relevant to analysing potentially exculpatory evidence.

83. In *Karadžić*, the Appeals Chamber of the IRMCT rejected Karadžić's contentions 'that the Trial Chamber erred in determining that his, orders were

87 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 1004.
88 *Abd-Al-Rahman* TC (Chapter 1, n 51) § 1005.

"not genuine", [and] were "politically motivated"'[89] and his assertion that 'his political motivation was irrelevant in assessing his intent'.[90] In assessing the orders issued by Karadžić prohibiting the targeting of civilians in Sarajevo, the Appeals Chamber determined that it had been reasonable for the Trial Chamber of the ICTY to consider the broader context.[91]

84. Because the orders Karadžić had given were issued in a context where he was 'pressured by the international community or threatened with air strikes' the Appeals Chamber determined that Karadžić had failed to demonstrate the unreasonableness and error of the Trial Chamber in their interpretation of his orders.[92] It therefore agreed with the conclusion at trial that, 'orders prohibiting the targeting of civilians did not indicate that Karadžić disapproved of the shelling and sniping of Sarajevo, but rather that they were conducted at times inconvenient to him'.[93]

85. Further, the Appeals Chamber noted 'the fact that his orders to stop the sniping and shelling in Sarajevo and to respect the laws of war had no practical effect on the situation on the ground as they were never followed up by any meaningful investigation and/or punishment for those who disobeyed them'.[94] As such, the Appeals Chamber upheld the Trial Chamber's use of his orders in their assessment of Karadžić's intent 'to commit murder, torture, and unlawful attacks on civilians in relation to the Sarajevo JCE (Joint Criminal Enterprise)'.[95]

2.5 Case Vignette: Mladić IRMCT Appeals Judgement at the ICTY

86. In the *Mladić* appeals case, which closely mirrored *Karadžić* in many respects, the defence asserted that the Trial Chamber had erred in its assessment of his *mens rea* in relation to furthering the overarching JCE by failing to adequately consider the orders he gave to subordinates to respect international law which he claimed contradicted a finding of criminal intent.[96] Defence Counsel argued that 'the Trial Chamber failed to give sufficient weight to the military context and contents of legitimate military orders he issued

89 *Prosecutor v Radovan Karadžić* (Judgement) MICT-13-55-A (20 March 2019) § 542.
90 *Karadžić* AC (Chapter 1, n 89) § 543.
91 *Karadžić* AC (Chapter 1, n 89) § 542.
92 *Karadžić* AC (Chapter 1, n 89) § 542.
93 *Karadžić* AC (Chapter 1, n 89) § 542.
94 *Karadžić* AC (Chapter 1, n 89) § 544.
95 *Karadžić* AC (Chapter 1, n 89) § 545.
96 *Prosecutor v Ratko Mladić* (Judgement) MICT-13-56-A (08 June 2021) § 249.

in Srebrenica',[97] and argued that his orders included 'genuine warnings in his orders for VRS soldiers to respect the Geneva Conventions' and 'direct orders to observe ceasefire warnings'.[98]

87. The Trial Chamber, in its assessment of Mladić's orders, had rejected the genuine nature of these orders and found them not reflective of his true intent or state of mind 'as they were inconsistent with his other conduct, and directly contradicted by his other contemporaneous statements'[99] and were 'contradicted by what happened on the ground'.[100] The Appeals Chamber ultimately agreed with the Trial Chamber that his orders to follow the Geneva Conventions were not genuine, based on the extensive assessment of evidence to contextualise the orders. This evidence included contemporaneous statements in which he had used 'derogatory terms to refer to Bosnian Muslims and Bosnian Croats [and] made reference to historical crimes committed against Bosnian Serbs' but also earlier statements made by Mladić in Croatia in 1991 and in Bosnia in 1992 'indicating intent not to respect the laws of war'.[101] The ICTY Appeals Chamber took into account events unfolding on the ground in its assessment of the orders given by the defendant, in order to contextualise the language of the order.[102] Further, on Mladić's assertion that his orders were legitimate within the military context of Srebrenica (where genocide subsequently occurred), the Appeals Chamber recalled that 'an accused's contribution to a joint criminal enterprise need not be in and of itself criminal, as long as he or she performs acts that in some way contribute to the furtherance of the common purpose'[103] and thus took the view that 'whether Mladić's orders were legitimate in the military context is not relevant to determining his significant contribution to the common purpose'.[104] Rather what matters is that 'these orders were so instrumental to the commission of the crimes that without them the crimes would not have been committed as they were'.[105]

88. The Appeals and Trial Chambers' assessment of Mladić's orders show that the context of the speech acts is relevant to interpreting both the meaning and legal significance of statements alleged to have contributed to atrocity

97 *Mladić* AC (Chapter 1, n 96) § 390.
98 *Mladić* AC (Chapter 1, n 96) § 249.
99 *Mladić* AC (Chapter 1, n 96) § 259.
100 *Mladić* AC (Chapter 1, n 96) § 257.
101 *Mladić* AC (Chapter 1, n 96) §§ 249–250, footnote 883.
102 *Mladić* AC (Chapter 1, n 96) § 396.
103 *Mladić* AC (Chapter 1, n 96) § 393.
104 *Mladić* AC (Chapter 1, n 96) § 393.
105 *Mladić* AC (Chapter 1, n 96) § 392. See also See *Prosecutor v Ratko Mladić* (Judgement) IT-09-92-T (22 November 2017) §§ 5097–5098.

crimes. However, reviewing the context does not displace a specific analysis of the contents and legal impact of each order. The fact there was an ongoing military conflict did not automatically insulate Mladić's statements from a determination of criminal responsibility. At the same time, the Chambers considered the context – particularly the commission of grave and large-scale crimes on the ground – to undermine an interpretation that the orders were a genuine instruction to subordinates to adhere to international humanitarian law. The generalisable implication is that speech acts do not need to be criminal on their face to attract liability. Instead, the question will be whether or not the statements (in this case orders) made a significant contribution to the crimes (or to the common criminal plan pursuant to which the crimes were committed) as a factual matter in the circumstances of the case.

89. Interpreting orders will take into account that those issuing them occupy a position of authority, by definition. With this authority comes greater inherent responsibility (and potential criminal liability) for the statement maker, due to the heightened implications of their words, particularly when framed as orders. These considerations will apply with equal force in the digital era, wherein the circumstances of the case will provide critical interpretive guidance to assess statements of the accused. Nonetheless, the emergence of social media as an increasingly ubiquitous phenomenon provides an additional source of interpretive material for international criminal tribunals. Posts on publicly accessible platforms issued contemporaneously with orders for operations on the ground will indicate how those orders are being explained externally, as elaborated on in the following sections on disinformation and social media.

CHAPTER 2

Disinformation and Speech Crimes

> A large number of the ordinary members of the German nation would never have participated in or tolerated the atrocities committed throughout Europe if they had not been conditioned to barbarous convictions and misconceptions by the constant grinding of the Nazi propaganda machine.[1]

90. Although coined in the 1920s in the Soviet Union in a special disinformation office,[2] the notion of disinformation has received increasing attention on the international stage in recent years. The Covid-19 pandemic, the Trump Administrations in the United States, and frequent election interference around the world in the last decade have all been accompanied by allegations about the critical role of disinformation campaigns. Both state actors (or government sponsored) and non-State actors may be behind the spread of disinformation, and these actors can be internal or external. International criminal proceedings have also frequently seen the waging of parallel information 'wars', with the Russian full-scale invasion of Ukraine starting in February 2022 and Israel's response to the 7 October 2023 attack by Hamas being particularly fertile grounds for the spread of disinformation, including through Russian bot farms. In recent years, it has become increasingly clear that the links between disinformation and atrocity crimes require close attention. Despite this, there remain few instances in which the dissemination of disinformation has been directly criminalised in international legal proceedings.[3]

91. At the outset, it should be noted that international criminal law is not a panacea against the harmful effects of disinformation. Moreover, disinformation should not be criminalised in an overly broad manner.[4] Indeed, the

1 Captain DA Sprecher and Thomas J Dodd, 'Oral Presentation Before the International Military Tribunal of the Individual Responsibility of the Defendant Hans Fritzsche for Crimes Against the Peace, War Crimes, and Crimes Against Humanity' (International Military Tribunal, Prosecution Oral Presentation Against Hans Fritzsche, Nuremberg 1946), University of Connecticut: Archives & Special Collections at the Thomas J Dodd Research Center 20; Cited in Wilson (Introduction, n 8) 282.

2 Aristedes Mahairas and Mikhail Dvilyanski, 'Disinformation – Дезинформация (Dezinformatsiya)' (2018) 3.3 The Cyber Defense Review 21.

3 Mathias Holvoet, 'International Criminal Liability for Spreading Disinformation in the Context of Mass Atrocity' (2022) 20:1 Journal of International Criminal Justice 226.

4 The United Nations (UN) Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe (OSCE) Representative on Freedom

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-004
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

internationally recognised human right to freedom of expression and to impart information and ideas is not limited to only sharing 'correct' or verifiable statements.[5] Even the intentional and knowing spread of mistruths should not be subjected to penal sanction under international criminal law without showing that the disinformation contributed significantly to one or more atrocity crimes.

92. UN GA Resolution 76/227 on Countering disinformation for the promotion and protection of human rights and fundamental freedoms adopted in 2021 stresses that 'responses to the spread of disinformation must comply with international human rights law and the principles of legality, necessity, and proportionality' reaffirming that 'the need to ensure that efforts to counter disinformation promote and protect and do not violate individuals' freedom of expression and freedom to seek, receive and impart information'.[6] The importance of this balance and the protection of freedom of expression is also emphasised in the 2018 European Union Action Plan Against Disinformation.[7]

93. Against this backdrop, the following analysis of international criminal jurisprudence highlights the ways in which disinformation may contribute to the commission of atrocity crimes. This may be done by specifically seeking to cover up crimes and distract international attention away from them, or it may be part of a broader strategy to foster and support an environment which is conducive to violence against a targeted group.[8]

---

of the Media, the Organization of American States (OAS) Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples' Rights (ACHPR) Special Rapporteur on Freedom of Expression and Access to Information, 'Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda' (*OSCE* 03 March 2017) <https://www.osce.org/files/f/documents/6/8/302796.pdf>.

5 See *Salov v Ukraine* App no 65518/01 (ECtHR 6 September 2005) § 113 '[…] art 10 of the Convention [ECHR] as such does not prohibit discussion or dissemination of information received even if it is strongly suspected that this information might not be truthful. To suggest otherwise would deprive persons of the right to express their views and opinions about statements made in the mass media and would thus place an unreasonable restriction on the freedom of expression set forth in art 10 of the Convention.'

6 UNGA Res 76/227 'Countering Disinformation for the Promotion and Protection of Human Rights and Fundamental Freedoms' (24 December 2021) UN Doc A/RES/76/227.

7 European Commission (EC) High Representative of the Union for Foreign Affairs and Security Policy, 'Joint Communication to The European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of The Regions: Action Plan against Disinformation' (05 December 2018) 1.

8 Holvoet (Chapter 2, n 3) 226–227; See also Martin Innes, Diyana Dobreva and Helen Innes, 'Disinformation and Digital Influencing After Terrorism: Spoofing, Truthing and Social Proofing' (2021) 16:2 Contemporary Social Science 1, 1–15 (discussing the effects of disinformation campaigns in the aftermath of terrorist attacks).

94. As an overarching organisational point, it must be noted that a common area of confusion is the relationship between (a) the empirical finding that disinformation contributes to atrocity crimes, e.g. by creating an environment conducive to violence, and (b) the normative question of the criminalisation of disinformation. At the international level, disinformation (and even disinformation for war) is not per se criminal as a legal-normative matter, but as a factual matter courts have found disinformation to be relevant to the commission of crimes, in several respects.

95. In the 21st Century, the topic of disinformation has grown in importance alongside the growth of social media. The phenomenon of social media has allowed for rapid and frictionless dissemination of disinformation to a global audience. Concomitantly, this has increased attention the potential role of social media in the commission of international crimes in recent or currently ongoing conflicts. This role was confirmed, for example, in the 2018 UN HRC Report of the Independent International Fact-Finding Mission on Myanmar (IIMM) which stated that 'the role of social media is significant' and, specifically, that social media [Facebook] had been a 'useful instrument for those seeking to spread hate'.[9] The report further went on to state that 'the extent to which Facebook posts and messages have led to real-world discrimination and violence must be independently and thoroughly examined'.[10] The topic of disinformation is thus closely related to the function of social media in facilitating political or military leaders to incite genocide and other atrocity crimes. In Chapter 3, we explore the question of whether senior staff of social media companies that provide a platform to inciting speech could be considered complicit and thus held liable for incitement to genocide under international criminal law.[11]

96. At the global and regional level, there exist several instruments to address disinformation, including the Joint Declaration on Freedom of Expression and 'Fake News', Disinformation and Propaganda, of 3 March 2017.[12] The Declaration provides that the fundamental human right to impart information and ideas is not limited to 'correct' statements and that Prohibitions on disinformation may violate international standards, but that nonetheless

9 Human Rights Council (HRC), 'Report of the independent international fact-finding mission on Myanmar' (12 September 2018) A/HRC/39/64 § 74 (HRC IIMM Report).

10 HRC IIMM Report (Chapter 2, n 9) § 74.

11 See Chapter 3, 'Corporate Criminal Liability for International Crimes'.

12 UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression and ACHPR Special Rapporteur on Freedom of Expression and Access to Information (Chapter 2, n 4).

does not justify the dissemination of State disinformation which may be extremely harmful.[13] Additionally, UNGA Resolution A/RES/76/227 UN GA adopted on 24 December 2021 on Countering disinformation and promotion and protection of human rights and fundamental freedoms, reiterates many of these principles.[14] The EU Action Plan Against Disinformation (2018) sets out key parameters for addressing disinformation, including the four pillars of: (i) improving the capabilities of Union institutions to detect, analyse and expose disinformation; (ii) strengthening coordinated and joint responses to disinformation; (iii) mobilising private sector to tackle disinformation; (iv) raising awareness and improving societal resilience.[15] Nonetheless, these are so-called 'soft law' instruments,[16] in international law. Consequently, the normative space to address disinformation at the international level remains wide open.

1 Conceptualising Disinformation in International Trials

97. It is first important to clarify that although the term disinformation is often used alongside others – such as misinformation, hate speech, and propaganda – these terms are related but not interchangeable. There is no universally accepted legal definition of misinformation and disinformation. For conceptual clarity, the definition employed here follows the approach developed by the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. The definition put forward by the Special Rapporteur understands disinformation as 'false information that is disseminated intentionally to cause serious social harm'.[17] As such, the spread of disinformation is a purposeful act intended to manipulate or mislead its audience

13 UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression and ACHPR Special Rapporteur on Freedom of Expression and Access to Information (Chapter 2, n 4).

14 UNGA Res 76/227 (Chapter 2, n 6).

15 EC Action Plan against Disinformation (Chapter 2, n 7) 5.

16 On the notion of soft law, see Sumudu Atapattu, 'International environmental law and soft law: a new direction or a contradiction?', in Cecilia Bailliet (ed.) *Non-state actors, soft law and protective regimes: from the margins*, (Cambridge University Press 2012); Although not legally binding, soft law instruments can influence the interpretation of terms of treaties and other primary sources of international law.

17 UNGA, Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, 'Disinformation and freedom of opinion and expression' (13 April 2021) A/HRC/47/25 § 15.

and supports a larger plan or agenda.[18] In contrast, misinformation is defined as 'the dissemination of false information unknowingly'.[19] As such, the difference lies with the underlying intent to cause harm.

98. These terms have at times been somewhat conflated in international criminal proceedings. There are cases discussed throughout this section where the term used by the Chamber is being described and discussed aligning more accurately with what is understood as "disinformation" due to the deliberately harmful nature of the relevant act.[20]

99. In a context of armed conflict, the concepts of hate speech, propaganda and disinformation are also often closely interrelated. While they may overlap with one another under some circumstances this is not always the case. Although both propaganda and disinformation 'imply a degree of purposeful and systematic manipulation of information',[21] there is a conceptual distinction to be made. For example, there may be propaganda campaigns that consist of the deliberate dissemination of false information – in which case, such propaganda would be a type of disinformation. There may, however, also be propaganda which, although it may be biased and spread with the intention of causing harm or to manipulate the public, could consist of true information (or at least of falsehoods that were thought to be true at the time). Legally, propaganda for war must be prohibited under Article 20(1) of the ICCPR, which is reflective of customary international law and thereby is substantively binding on signatory States.

2 Disinformation and Propaganda as Addressed by International Courts

100. Cases with a central element of propaganda have featured in international criminal law since the early days of its emergence as an established legal regime. For example, the hate speech proceedings which were brought

18 Note that not all propaganda is disinformation. Moreover, there might also be disinformation that is used for purposes other than propaganda in its classic sense, such as to distract the security sector, public, or international community from other military operations or atrocities taking place.

19 Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (Chapter 2, n 17) § 15.

20 See for example the use of the term 'misinformation' in the *Mladić* Trial Chamber Judgement and Appeals Chamber Judgement to refer to deliberately misleading information. *Mladić* TC (Chapter 1, n 105) § 4502.

21 Naja Bentzen, 'Understanding Propaganda and Disinformation' (European Parliamentary Research Service, EPRS_ATA (November 2015) 571332.

against Julius Streicher, editor of a Nazi newspaper, and Hans Fritzsche, Head of the Nazi Radio Division, at the Nuremburg International Military Tribunal (IMT) and the trial of Nazi Press Chief Otto Dietrich at the subsequent United States Nuremburg Military Tribunals pursuant to Control Council Law No. 10.[22] Disinformation and propaganda have also featured in the jurisprudence of the ICTY[23] and the ICTR.[24]

101. Disinformation, including propaganda based on disinformation, may play a central enabling role in creating an environment conducive to the commission of atrocity crimes. In this context, disinformation campaigns can take several different forms, have different purposes, and accordingly may occur at different time points in relation to the commission of mass atrocity crimes. The case studies presented in the following sub-sections examine examples from international criminal law jurisprudence of disinformation campaigns disseminated prior to the commission of atrocity crimes, with an inciting effect, during their commission and with the purpose of furthering or fuelling their ongoing commission, and after the commission of atrocity crimes, for example to cover-up and deny that crimes took place.[25]

102. The subsections below discuss how the dissemination of disinformation, including through propaganda campaigns or through the deliberate concealment of atrocity crimes, could engage liability under international criminal law. They consider how disinformation could potentially constitute an inchoate speech crime in cases of incitement to genocide or participation in persecution (as a crime against humanity) and also take into account how disinformation campaigns may contribute to the commission of an offence, particularly through the mode of liability of aiding and abetting. Several important limitations are discussed, including the contemporaneousness and temporal link of disinformation campaigns compared to the commission of an atrocity crime, challenges establishing a causal link, and questions regarding

22 Gregory Gordon, 'The Forgotten Nuremberg Hate Speech Case: Otto Dietrich and the Future of Persecution Law' (2014) 75:3 Ohio State Law Journal 571, 576; Gregory Gordon, 'The Propaganda Prosecutions at Nuremberg: The Origin of Atrocity Speech Law and the Touchstone for Normative Evolution' (2017) 39 Loyola of Los Angeles International and Comparative Law Review, Special Edition: The Nuremberg Laws and the Nuremberg Trials 209, 231.

23 Michael G Kearney, 'Propaganda in the jurisprudence of the international criminal tribunal for the former Yugoslavia' in Predrag Dojčinović (ed) *Propaganda, War Crimes Trials and International Law from Speakers' Corner to War Crimes* (Routledge 2011) 231–253.

24 See the discussion of *Nahimana et al.* in Chapter 3 below, showing that statements were made which misleadingly labelled civilian Tutsis as threats to Hutus via RTLM and Kangura.

25 Holvoet (Chapter 2, n 3) 226–227.

the role played by a disinformation campaign and its ultimate impact depending on the audience targeted by the campaign.

3 Case Vignettes on Disinformation Campaigns and Atrocity Crimes

103. The question of how disinformation campaigns may contribute to the commission of atrocity crimes (or even constitute crimes per se) has featured in key international criminal law proceedings. As previously noted, this has in some cases been closely linked with overlapping discussions on the role of propaganda in armed conflict and as a potential contributing factor leading to the commission of atrocity crimes.

104. The following review of key international trials which discussed disinformation raises additional factors to consider when evaluating the potential contribution of disinformation to the commission of atrocity crimes. Factors which are highly likely to emerge in any such discussion include both the accused's knowledge of the false nature of the information they disseminate, closely linked to their intent and purpose in spreading the incorrect information, and the audience targeted by the disinformation campaign. These have proven crucial at trial. The sections below discuss cases where disinformation has proved crucial in an international trial, including; the acquittal of Nazi propagandist Hans Fritzsche, the decision not to confirm charges in the ICC cases *Mbarushimana* and *Ruto and Sang*, as well as in the convictions of Streicher in the post-World War II trials, and the cases of *Karadžić*, *Mladić* and *Gvero* at the ICTY.

105. On the basis of our review of this body of case law, we recommend that courts dealing with international crimes adopt a two-pronged test to evaluate and determine disinformation of atrocity crimes (namely, when the concealment or denial of crimes is not neutral but designed for a purpose) consisting of the following elements set out in the *Mbarushimana* Pre-Trial Chamber confirmation decision: (i) establishing that the accused has awareness of the crimes committed and knows that he is denying the truth; and (ii) proving that this denial of the truth is done in furtherance of a policy or common purpose.[26]

3.1 Case Vignette: Fritzsche at the IMT at Nuremburg, 1945

106. The fact that disinformation played a role at the IMT demonstrates a long-standing interest in international criminal tribunals regarding the

26 *Prosecutor v Callixte Mbarushimana*, (Decision on the Confirmation of Charges) ICC-01/04-01/10-465-Red (16 December 2011) § 312.

connection between misleading and untrue statements and the perpetration of atrocity crimes. Although modern international criminal law is constituted differently than Nuremberg, it remains firmly based on many of the core concepts and legal principles and formulations established during the post-World War Two trials.

107. The trial of Hans Fritzsche, the Head of the German Home Press Department at Joseph Goebbel's Ministry of Propaganda and host of a daily radio show, at the IMT at Nuremburg provides an important starting point for this analysis, despite the fact that he was acquitted on the grounds that Fritzsche's main goal was to rally public support for Adolf Hitler and the German war effort, and there was not enough evidence to prove his intent was to incite atrocities against a civilian population, and notably against Jews. Because of his role as a Nazi propagandist, Fritzsche was indicted for having 'the foregoing positions and his personal influence to disseminate and exploit the principle doctrines of Nazi conspirators set forth in Count One of the Indictment, and to advocate, encourage and incite the commission of the War Crimes set forth in Count Three of the Indictment and the Crimes against Humanity set forth in Count Four of the Indictment including, particularly, anti-Jewish measures and the ruthless exploitation of occupied territories'.[27]

108. In relation to the charge of crimes against peace,[28] it was noted that disinformation featured as a frequent element of Fritzsche's 'vigorous propaganda campaigns carried out before each major act of aggression'.[29] In his affidavit, Fritzsche describes the increasingly propagandist treatment of Poland in the German press in the Spring and throughout the Summer of 1939 ahead of Germany's invasion of Poland when German newspapers received instructions to publish accounts of the mistreatment and cruelty directed towards the German minority in Poland, provocation of border incidents by Poland and a Polish desire for conquest (*Eroberungsgelueste*).[30]

109. Notably, the IMT found that although Fritzsche 'held daily press conferences to deliver the directives of the Propaganda Ministry' to German newspapers, he was doing so in a subordinate position meaning that he 'had no

27 *Trial of the Major War Criminals Before the International Military Tribunal* (*IMT*) *at Nuremberg* (Nuremberg 14 November 1945 to 1 October 1946) Volume 1, 79.

28 Crimes against peace in the IMT statute are a forerunner to the crime of aggression in modern day international criminal law. See Carrie McDougall, 'The Crimes Against Peace Precedent' in Claus Kreß and Stefan Barriga (eds), *The Crime of Aggression – A Commentary* (Cambridge University Press 2017) Chapter 2, 49–112.

29 *IMT Vol. 1* (Chapter 2, n 27) 337.

30 *Trial of the Major War Criminals Before the International Military Tribunal* (*IMT*) *at Nuremberg* – Erklärung des Hans Fritzsche (Nuremberg 14 November 1945 to 1 October 1946) Volume 32, 320–321.

control over the formulation of these propaganda policies' and was 'merely a conduit to the press of the instructions handed him by Dietrich' (himself a subordinate to Goebbels).[31] For the charge of crimes against peace, it was noted that Fritzsche had never achieved 'sufficient stature to attend the planning conference which led to aggressive war' and as such 'his activities cannot be said to be those which fall within the definition of the common plan to wage aggressive war'.[32] This can be seen as reflecting the leadership elements of aggression as well as the need to show a meaningful contribution to the crime that was committed to be a common plan participant.

110. As the IMT framework did not plainly include complicity as a mode of liability in the list of four crimes that could be charged, Fritzsche had to be charged as a participant in the common plan, or conspiracy to wage a war of aggression.[33] Pre-trial briefs provide some insight into how the judges viewed Fritzsche's role as 'a principal conspirator in abetting aggressive wars', who through his actions created 'the requisite psychological and political conditions for aggressive wars' among the German people.[34] There was, then, a gap between the relevant mode of liability available under the IMT framework to charge Fritzsche and the IMT judges' perception of his contribution to crimes against peace.[35]

111. In his dissent, Soviet Judge Nikitchenko disagreed with the Tribunal's reasoning that Fritzsche had 'not reached in Germany the official position making him responsible for the criminal actions of the Hitler regime and that his own personal activity in this respect cannot be considered criminal'.[36] Nikitchenko considered propaganda, referencing particularly 'the false presentation of facts'[37] (i.e. disinformation used for the purpose of propaganda) to be one of the most important weapons of the Nazi regime stating that 'the dissemination of provocative lies and the systematic deception of public opinion

31 *IMT Vol. 1* (Chapter 2, n 27) 336.

32 *IMT Vol. 1* (Chapter 2, n 27) 337.

33 Holvoet (Chapter 2, n 3) 231.

34 Richard A Wilson, *Incitement on Trial: Prosecuting International Speech Crimes* (Cambridge University Press 2017) 28.

35 For an interesting discussion on the potential for accountability under art 25 of the Rome Statute framework on aiding and abetting as a mode of liability applied to perpetrators like Fritzsche regarding the crime of aggression see Holvoet (Chapter 2, n 3) 231.

36 *IMT Vol. 1* (Chapter 2, n 27) Dissenting Opinion of the Soviet Member of the International Military Tribunal, III. The Unfounded Acquittal of Defendant Fritzsche 350.

37 *IMT Vol. 1* (Chapter 2, n 27) Dissenting Opinion of the Soviet Member of the International Military Tribunal, III. The Unfounded Acquittal of Defendant Fritzsche 351.

were as necessary to the Hitlerites for the realisation of their plans as were the production of armaments and the drafting of military plans'.[38]

112. With respect to the charges of war crimes and crimes against humanity, the Prosecution asserted that 'Fritzsche incited and encouraged 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 under Counts Three and Four'.[39] The IMT judges found, that 'his position and official duties were not sufficiently important, however, to infer that he took part in originating or formulating propaganda campaigns'.[40] Furthermore, while it was noted that 'evidence from his speeches show definite anti-Semitism on his part' these speeches did not directly urge 'persecution or extermination of Jews' and that 'there is no evidence that he [Fritzsche] was aware of their extermination in the East'.[41] He added that '[w]ithout propaganda, founded on the total eclipse of the freedom of press and of speech, it would not have been possible for German fascism to realise its aggressive intentions, to lay the groundwork and then to put to practice the War Crimes and the Crimes against Humanity.'[42]

113. For his role in the commission of war crimes and crimes against humanity, Fritzsche was charged under the mode of liability of instigation. This again highlights the gap between the relevant modes of liability available under the IMT framework and the judges' perception of Fritzsche's contribution. As highlighted by Holvoet, the IMT judges viewed Fritzsche's disinformation propaganda campaigns as being 'insufficiently direct' to constitute instigation to the charged war crimes and crimes against humanity.[43] The Judgement concluded instead that Fritzsche's aim had been to condition the German people and arouse popular support for Hitler's war effort. This conclusion could indicate that even at the time of the IMT, 'the judges considered aiding and abetting as a more apt way to conceptualise how a non-military propagandist like Fritzsche could have contributed to the crimes'.[44] It is also important to note that despite his acquittal at Nuremberg, Fritzsche was subsequently

38 *IMT Vol. 1* (Chapter 2, n 27) Dissenting Opinion of the Soviet Member of the International Military Tribunal, III. The Unfounded Acquittal of Defendant Fritzsche 351.

39 *IMT Vol. 1* (Chapter 2, n 27) 337–338.

40 *IMT Vol. 1* (Chapter 2, n 27) 338.

41 *IMT Vol. 1* (Chapter 2, n 27) 338. This later proved to be false, see Holvoet (Chapter 2, n 3) 234.

42 *IMT Vol. 1* (Chapter 2, n 27) Dissenting Opinion of the Soviet Member of the International Military Tribunal, III. The Unfounded Acquittal of Defendant Fritzsche, 351.

43 Holvoet (Chapter 2, n 3) 233.

44 Holvoet (Chapter 2, n 3) 233.

re-arrested and convicted by a West German court for being a major Nazi war leader, resulting in a sentence of 9 years' imprisonment.[45]

3.2 Case Vignette: Julius Streicher at the Nuremberg Tribunal

114. In contrast to the *Fritzsche* ruling at the IMT, Julius Streicher was convicted and hanged at Nuremberg for crimes against humanity. Streicher's case was exceptional in that he was the only defendant convicted and executed at Nuremberg solely on Count Four of Crimes Against Humanity. The prosecution theory of the case was aiding and abetting, although this was not made explicit until after the trial.[46] The Nuremberg tribunal convicted Julius Streicher, on Count 4 (crimes against humanity), holding that '[i]n his speeches and articles … [he] incited the German People to active persecution'.[47] It concluded that 'Streicher's incitement to murder and extermination, at the time when Jews in the East were being killed under the most horrible conditions, clearly constitutes persecution on political and racial grounds in connection with War Crimes, as defined by the Charter, and constitutes a Crime against Humanity'.[48] The primary evidence against Streicher was an article he wrote in the May 1939 issue of the pro-Nazi publication, *Der Stürmer*, which stated that '[t]he Jews in Russia must be killed. They must be exterminated root and branch',[49] and the fact that his articles in the pro-Nazi publication *Der Stürmer* 'termed the Jew as a germ and a pest, not a human being, but "a parasite, an enemy, an evildoer, a disseminator of diseases who must be destroyed in the interest of mankind"'.[50] In the way it recognised Streicher's contribution to the genocidal programme of the Nazis, the Tribunal's decision runs counter to its finding in *Fritzsche* that propaganda is only of minor importance in the commission of crimes against humanity. Unfortunately, we cannot conclude that the IMT provided a clear message on how international courts should approach the question of the potential criminal liability of propagandists and purveyors of disinformation.

---

45 Gordon (Chapter 2, n 22); Regarding this, The New York Times reported that: 'The distortions of truth for which he was responsible as chief of the German press section, it found, were not merely those necessitated by war conditions, but served to disseminate Nazi ideology. The tribunal also ruled that as an author and as 'spokesman of the Government to the German people' on the radio he showed complete agreement with the basic Nazi principles.' See Dana Adams Schmidt, 'Germans Give Fritzsche 9 Years; Hitler Photographer Receives 10' (*New York Times* 02 February 1947) <https://www.nytimes.com/1947/02/01/archives/germans-give-fritzsche-9-years-hitler-photographer-receives-10.html>.

46 Wilson (Chapter 2, n 34) 28.

47 Wilson and Gillett (Key Guidelines, n 1) § 121, citing *IMT Vol 1.* (Chapter 2, n 27) 302.

48 *IMT Vol. 1* (Chapter 2, n 27) 304.

49 *IMT Vol. 1* (Chapter 2, n 27) 302–303.

50 See Wilson and Gillett (Key Guidelines, n 1) § 122, citing *IMT Vol 1.* (Chapter 2, n 27) 303.

3.3 Case Vignette: Gvero at the ICTY

115. Questions about liability for the spread of disinformation also arose in the trial of Bosnian Serb General Milan Gvero. Gvero was charged as one of the accused in the *Popović et al.* at the ICTY, which concerned the displacement and large-scale killing of Bosnian Muslims. Gvero was not convicted for the killing operation, but instead just for his role in preparing information that contributed to the displacement of around 40,000 Bosnian Muslims from the Srebrenica and Žepa enclaves in 1995. Specifically, the Trial Chamber held that, in his role as Assistant Commander for Morale, Legal and Religious Affairs, Gvero was tasked with 'the dissemination of information and propaganda for the troops in support of the aims of the war, in the preparation for and during the course of combat operations'.[51] Gvero was convicted for the crimes of humanity of persecution and inhumane acts (forcible transfer) under Articles 5(h) and 5(i) of the ICTY Statute. In relation to these crimes, Gvero was found responsible pursuant to Article 7(1) of the ICTY Statute 'for planning, instigating, ordering, committing, and otherwise aiding and abetting' their commission including through the dissemination of false information.[52]

116. By knowingly disseminating false and misleading information to the international community during the ongoing commission of atrocity crimes in the Srebrenica and Žepa enclaves, Gvero was accused of having committed acts in furtherance of the Joint Criminal Enterprise (JCE).[53] In his indictment, the prosecution stated that Gvero 'released to the public media a false statement concerning the attacks on the enclaves in order to assist in the take-down of the Srebrenica enclave'.[54] The indictment further alleged that Gvero 'assisted in the attack on Srebrenica, knowing that one of the main objectives was to force the Muslim population to leave Srebrenica, by lying to [United Nations Protection Force] UNPROFOR about Muslim attacks, in particular on [United Nations Office for Project Services] UN OPS, and [Army of Republika Srpska] VRS intentions and actions with respect to the enclave'.[55]

117. The Chamber found that Gvero 'was entrusted with key functions related to external propaganda and interaction with international organisations, aimed at supporting the plan to forcibly transfer the populations of

51 *Prosecutor v Vujadin Popović, Ljubiša Beara, Drago Nikolić, Ljubomir Borovčanin, Radivoje Miletić, Milan Gvero, Vinko Pandurević* (Judgement) IT-05-88-T (10 June 2010) § 1752; also referred to as *Popović et al.* throughout.

52 *Popović et al.* (Chapter 2, n 51) § 1736, see also Gvero's role in disseminating false information see § 1814–1815 and conviction in § 2109.

53 *Popović et al.* (Revised Second Consolidated Indictment) IT-05-88-PT (14 June 2006) § 76.

54 *Popović et al.* Second Indictment (Chapter 2, n 53) § 76(a)(i).

55 *Popović et al.* TC (Chapter 2, n 51) § 76(b)(i).

Srebrenica and Žepa'.[56] In the section entitled, 'War Propaganda, Misleading Information and Threats' of the Judgement *Popović et al.*, the Trial Chamber considered Gvero's role 'in disseminating false information to the media and directly to international organisations'[57] alongside threats and pressure exerted on UNPROFOR Commander Nicolai in an effort to stop (North Atlantic Treaty Organization) NATO bombings, and how these statements by the defendant were interrelated.[58]

118. Regarding false information released to the international community, the Trial Chamber found that when Gvero had issued a statement to the media concerning the attack on the Srebrenica enclave 'he explained that the VRS' combat activity was directed towards neutralising Muslim terrorists and was not directed against civilians or members of UNPROFOR'.[59] While the Trial Chamber acknowledged that 'the release of false information to the media and international authorities does not constitute a criminal act' they evaluated the intent and purpose behind Gvero's press release.[60]

119. In the conclusion of his *communiqué* Gvero stated that 'there is no reason for the media and foreigners to get involved in the Muslim war propaganda'.[61] Considering Gvero's knowledge of the imminent VRS plan to 'take-over the Srebrenica enclave and to forcibly remove the civilian population, and the action which had been taken to implement it' the Trial Chamber found that this was 'unquestionably a misleading press release' the purpose of which was 'not an innocent one'.[62] In particular, the Trial Chamber found that the only reasonable inference as to the purpose of Gvero's press release was 'that it was intended to mislead [...] the international authorities concerned with protecting the enclave, with a view to delaying any action on their part which might thwart the VRS' military efforts'.[63]

120. When NATO initiated air strikes against the advancing VRS forces the following day, Gvero – in an effort to stop the bombing – called the UNPROFOR headquarters in Sarajevo and spoke with General Nicolai.[64] At this time, Gvero

56 *Popović et al.* TC (Chapter 2, n 51) § 1813.

57 *Popović et al.* TC (Chapter 2, n 51) § 1812; See also *Popović et al.* Second Indictment (Chapter 2, n 53) § 76(a)(i) and 76(b)(i).

58 *Popović et al.* TC (Chapter 2, n 51) § 1812; See also *Popović et al.* Second Indictment (Chapter 2, n 53) § 76(c)(i).

59 *Popović et al.* TC (Chapter 2, n 51) § 1814.

60 *Popović et al.* TC (Chapter 2, n 51) § 1815.

61 *Popović et al.* TC (Chapter 2, n 51) § 1768, footnote 5324 citing Ex. P02753, 'Srebrenica – The Muslim War Trump Card', statement by Gvero (10 July 1995) 2.

62 *Popović et al.* TC (Chapter 2, n 51) § 1815.

63 *Popović et al.* TC (Chapter 2, n 51) § 1815.

64 The Dutch brigadier who was Chief of Staff of UNPROFOR at Sarajevo during the fall of Srebrenica and acted as liaison officer for the Dutch government.

had knowledge that the VRS had taken over Srebrenica and that Bosnian Muslims had fled the enclave and gathered at the DutchBat compound in Potočari.[65] Yet during this conversation with Nicolai, he adopted the same 'tactic of falsely asserting that this was only the VRS responding to attacks'[66] by threatening that unless NATO air support was immediately stopped, Nicolai 'would be held responsible for all further developments and the destiny of his men and the civilian population in Srebrenica'.[67] While the Trial Chamber acknowledged that although 'no overt threatening language was used nor were possible reprisals against the people gathered in Potočari specifically referenced' it nevertheless concluded that Gvero's call was both intended and understood as a threat, given its timing in relation to the unfolding situation on the ground and the related efforts undertaken by the VRS to stop the NATO bombing.[68]

121. Alongside their recognition that the release of false information to the media did not in itself constitute a criminal act, the Trial Chamber also noted that they were 'unable to conclude that Gvero's discussion with Nicolai constituted a determinative factor in the NATO decision to suspend air support'.[69] It was, however, satisfied that stopping protective action by UNPROFOR and NATO in favour of the Srebrenica enclave – which essentially constituted 'the only means by which the VRS' advance towards and capture of the enclave could be halted' – was precisely Gvero's aim.[70] As such, the Trial Chamber found that 'on the basis of these acts taken in combination [...] Gvero played a limited but important role in supporting the VRS' military action, which was essential to the success of the plan to forcibly transfer'.[71] Further, the Trial Chamber was satisfied 'that by disseminating false information and issuing a serious threat, whether effective or not in the end, Gvero made a contribution to the JCE which by its nature cannot be classified as other than significant'.[72]

122. Elements of knowledge and shared intent were essential to finding Gvero criminally responsible for crimes against humanity under Article 5 of the ICTY Statute. Gvero's efforts to delay and block protective international intervention were considered against his detailed knowledge of the strategic

65 *Popović et al.* TC (Chapter 2, n 51) § 1816, see also for context § 253–255, 263.
66 *Popović et al.* TC (Chapter 2, n 51) § 1816.
67 *Popović et al.* TC (Chapter 2, n 51) § 1770, footnote 5328 citing Ex. P02906, 'Notes of a telephone conversation between Nicolai and Gvero' (11 July 1995, 16:15 hours).
68 *Popović et al.* TC (Chapter 2, n 51) § 1816.
69 *Popović et al.* TC (Chapter 2, n 51) § 1818.
70 *Popović et al.* TC (Chapter 2, n 51) § 1818.
71 *Popović et al.* TC (Chapter 2, n 51) § 1820.
72 *Popović et al.* TC (Chapter 2, n 51) § 1820.

aim to forcibly remove the Bosnian Muslim population from the Srebrenica and Žepa enclaves.[73] Further, in intercepted conversations Gvero was found to have said that 'everything is going according to plan, don't worry' thereby evidencing not only that he met the knowledge requirement but also that he shared intent.[74]

123. As such, the Trial Chamber determined that Gvero was 'guilty of inhumane acts (forcible transfer) as a crime against humanity punishable under Article 5(i) of the Statute'[75] due to his role in the 'commission of the forcible transfer of the Bosnian Muslims from Srebrenica and Žepa through his participation in the JCE to Forcibly Remove'.[76] The Trial Chamber found that Gvero's direct involvement in any of the logistical aspects of carrying out the forcible transfer operations could not be demonstrated.[77] His actions of intervening with international organisations and entities through the release of false information to the media and the threats he made to General Nicolai 'evidence that he was involved in the JCE in the planning and development stages and subsequently' but the Trial Chamber was not satisfied that 'the only reasonable inference is that Gvero, through his knowledge and through the level of his involvement in the forcible transfer operation would foresee that "opportunistic" killings would be perpetrated by members of the VRS in the course of the forcible transfer'.[78]

124. It is important to note that Gvero received a relatively lenient sentence of 5 years of imprisonment, as the Trial Chamber's found that his contributions to the JCE were 'not numerous' and 'not decisive for the implementation of the common purpose'.[79] Further, there was a lack of evidence regarding his involvement in decision-making on military actions related to the plan.[80]

125. In the *Gvero* case, disinformation consisted of false statements and press releases made to the international community and media to cover-up the commission of ongoing atrocity crimes and VRS attacks in the Srebrenica and Žepa enclaves. The treatment of disinformation in the *Gvero* trial raises several important points. Notably, the Trial Chamber affirmed that although the release of false information to the media and international community in and of itself does not constitute a criminal act, the purpose and intent behind

---

73 *Popović et al.* TC (Chapter 2, n 51) §§ 1822–1824.
74 *Popović et al.* TC (Chapter 2, n 51) § 1822.
75 *Popović et al.* TC (Chapter 2, n 51) § 1826.
76 *Popović et al.* TC (Chapter 2, n 51) § 1825.
77 *Popović et al.* TC (Chapter 2, n 51) § 1830.
78 *Popović et al.* TC (Chapter 2, n 51) § 1830.
79 *Popović et al.* TC (Chapter 2, n 51) § 2203.
80 *Popović et al.* TC (Chapter 2, n 51) § 2203.

its dissemination and its contribution to the JCE could see such conduct result in liability under international criminal law. As such, determining both the underlying intent and ultimately the contribution of disinformation in this case required evaluating the statements and press releases made by Gvero according to several aspects; the context within which they were made, the audience targeted by their dissemination, and their effectiveness.

126. At trial, the statements and press releases made by Gvero were evaluated against the context within which they were made and alongside the ongoing attacks and atrocities which were unfolding in the Srebrenica and Žepa enclaves. In particular, the Trial Chamber highlighted Gvero's knowledge of ongoing VRS attacks and its imminent plan to take-over Srebrenica and forcibly transfer its Muslim civilian population. This demonstrated the deliberately misleading nature of Gvero's statements and their purpose in delaying any course of action by international authorities that might thwart the VRS efforts.

127. Related to this are the considerations around the audience targeted by the dissemination of disinformation. It is therefore important to reiterate that protective actions undertaken by UNPROFOR and NATO constituted the sole means of halting, or at least delaying, the VRS advance towards and capture of Srebrenica. Given Gvero's knowledge of the plan to forcibly transfer the Bosnian Muslim population from the enclaves, the Trial Chamber was satisfied that his aim in disseminating disinformation was to stop any protective action. The target audience of Gvero's statements were not VRS troops and his aim in making them was not to directly incite or provoke further attacks. Rather, in targeting his disinformation campaign to the international community as an audience, Gvero sought to create conditions which allowed the already ongoing commission of crimes by the VRS to continue unabated and for further plans to be put in motion without intervention.

128. Finally, it is also noteworthy that in discussing the significance of Gvero's contribution to the JCE through the spread of disinformation, the Trial Chamber was unable to conclude that Gvero's statements or discussion with General Nicolai constituted a determinative factor in the decision to suspend NATO air support. Despite the inability to clearly determine the effectiveness of the disinformation disseminated by Gvero, the Trial Chamber nevertheless classified Gvero's contribution to the JCE as being significant, if limited. This evaluation of his contribution thus ascribed less importance to the actual effectiveness of Gvero's disinformation campaigns and rather determined its significance on the basis of his knowledge of ongoing atrocity crimes and his intent behind attempting to cover up said crimes in his dealings with the international community.

3.4 Case Vignette: Mladić at the ICTY

129. The trial of Ratko Mladić at the ICTY also dealt with the topic of deliberately misleading statements made to representatives of the international community, non-governmental organisations (NGOs), the media and the public and how disinformation facilitated the commission of crimes.[81] Specifically, this arose regarding the dissemination to the public of propaganda and disinformation[82] 'in order to cover up our true intentions and forces'[83] and false information provided by Mladić about the conditions of Prisoners of War (POWs) held in Serb detention camps.[84] The role of disinformation is important in the context of a litigated case, as it demonstrates that even this relatively indirect type of activity can have a significant impact on the crimes committed on the ground, by creating a permissive atmosphere in which crimes proliferate, especially when directed at vulnerable individuals or populations.

130. Regarding Mladić's contribution to the commission of atrocity crimes through the spread of disinformation, the Trial Chamber discussed his participation in establishing the procedures for propaganda to be disseminated. According to Gvero, Mladić issued instructions to all VRS commands about how to inform the local and international public on VRS activities on the basis of Mladić's orders which included instructions for preparing 'appropriate reports from the field'.[85] The Trial Chamber found that Mladić had not only participated in establishing the propaganda machinery but that he used said machinery to disseminate information released to the public in order to conceal the VRS forces' real intent and justify the war as necessary.[86]

---

81 See *Mladić* TC (Chapter 1, n 105) Chapter 9.3.8 Disseminating, encouraging and/or facilitating the dissemination of propaganda to Bosnian Serbs intended to engender in Bosnian Serbs fear and hatred of Bosnian Muslims and Bosnian Croats or to otherwise win support for and participation in achieving the objective of the joint criminal enterprise §§ 4480–4501; and Chapter 9.3.9 Engaging in, supporting and/or facilitating efforts to deny or to provide misleading information about crimes against Bosnian Muslims and Bosnian Croats and about the role that Serb Forces had played in those crimes to representatives of the international community, non-governmental organisations, the media and the public, thereby facilitating the commission of crimes §§ 4502–4513.

82 It should be noted here that although the Trial Chamber employed the term 'misinformation' the term 'disinformation' more accurately captures what is being discussed to the deliberately misleading and malicious nature of the statements made. The term disinformation is therefore used throughout this section, with the exception of direct quotes which may utilise misinformation in the meaning of disinformation.

83 *Mladić* TC (Chapter 1, n 105) § 4481.

84 *Mladić* TC (Chapter 1, n 105) § 4512.

85 *Mladić* TC (Chapter 1, n 105) § 4495.

86 *Mladić* TC (Chapter 1, n 105) § 4498.

131. An insider witness testified that Mladić 'did not tolerate lies' and that he 'requested that the public be informed objectively and truthfully about the situation in the territory and combat zone activities'[87] and Mladić himself had made statements in an interview with the magazine *Srpska Vojska* that he hoped 'the public would be informed objectively and truthfully about VRS combat activities and achievements'.[88] However, the Trial Chamber found that this evidence was inconsistent with the aforementioned orders given by Mladić to the Sector for Morale, Religious, and Legal Affairs of the VRS regarding the dissemination of false information in order to deceive the public about the 'true intentions and forces of the VRS' for the purpose of achieving the desired propaganda effects.[89] The Trial Chamber understood the desired propaganda effects to refer to a mobilisation of people of the Serb ethnicity, including by engendering fear and hatred of Bosnian Muslims and Bosnian Croats.[90] This was done on the basis of releasing false information about VRS intentions and activities, through the evocation of alleged historical narratives of Serb victimisation,[91] and the use of derogatory language to refer to Bosnian Muslims and Bosnian Croats.[92] The Trial Chamber thus found that Mladić, alongside his other activities, had participated in and significantly contributed to the JCE by having 'introduced and maintained a controlled and centralised system of spreading propaganda'.[93]

132. In addition to Mladić's control over the VRS propaganda machine to spread false information regarding VRS intent and activities to the local and international public, the topic of the provision of disinformation and its link to the commission of atrocity crimes also arose in relation to other acts. With regards to Mladić's contribution to the JCE through efforts to deny or provide misleading information about crimes against Bosnian Muslims and Bosnian Croats, thereby facilitating the commission of crimes, the Trial Chamber also discussed disinformation about VRS attacks on Sarajevo and related to the conditions inside Serb detention camps.

133. With regards to VRS attacks, the Trial Chamber considered evidence of a meeting in Belgrade on 10 June 1992, Mladić and other representatives including Karadžić and Krajišnik 'during which they, *inter alia*, discussed

87 *Mladić* TC (Chapter 1, n 105) § 4496.
88 *Mladić* TC (Chapter 1, n 105) § 4497.
89 *Mladić* TC (Chapter 1, n 105) § 4498.
90 *Mladić* TC (Chapter 1, n 105) § 4498.
91 *Mladić* TC (Chapter 1, n 105) § 4499.
92 *Mladić* TC (Chapter 1, n 105) § 4499.
93 *Mladić* TC (Chapter 1, n 105) § 4498, 4644.

the provision of misinformation regarding attacks on Sarajevo'.[94] Anthony Banbury, a UNPROFOR Civil Affairs Officer based in Sarajevo between 1994 and May 1995 who was later the Assistant to the UN Special Representative of the Secretary-General in Bosnia-Herzegovina provided testimony that 'that it was a widely held view among UNPROFOR officials dealing with Mladić that he very often misrepresented the truth to them'.[95]

134. The Trial Chamber also discussed Mladić's alleged contribution through his provision of disinformation provided to UNPROFOR officials, International Committee of the Red Cross (ICRC) representatives, and international media on the conditions inside Serb detention and POW camps. This included false statements made by Mladić when he claimed that treatment of prisoners inside the camps followed the Geneva Conventions[96] or that 'more than 5,000 POWs had been released'.[97] John Wilson, the chief United Nations Military Observer (UNMO) for UNPROFOR deployed to Sarajevo from 22 March to 24 June 1992, stated that at a meeting in Geneva that in December 1992 or January 1993 he spoke with Mladić about a photo circulating in the media of a malnourished man detained in a Serb camp.[98] Wilson testified that 'Mladić responded that the photographer had picked a particularly skinny man and that the man's condition did not accurately represent the treatment of persons detained by the Serbs, who looked after their detainees and fed them properly'.[99] When questioned on media coverage of allegations concerning the camps, journalist Edward Vulliamy[100] recalled that 'Mladić stated that the photographs from the camps were forgeries and montages, or else they were photographs showing Serbian prisoners in Muslim camps'.[101] Further, in an April 1993 meeting with ICRC representatives, Mladić made false statements regarding the release of POWs and claimed that ICRC efforts in Serb detention camps 'had been exploited by the media, and that [...] foreign media had raised tensions and turned the international community against the Serbs'.[102]

135. Evidence reviewed by the Trial Chamber directly contradicted the claims made by Mladić that the treatment of detainees was in accordance

94 *Mladić* TC (Chapter 1, n 105) § 4503.
95 *Mladić* TC (Chapter 1, n 105) § 4503.
96 *Mladić* TC (Chapter 1, n 105) § 4506.
97 *Mladić* TC (Chapter 1, n 105) § 4505.
98 *Mladić* TC (Chapter 1, n 105) § 4504.
99 *Mladić* TC (Chapter 1, n 105) § 4504.
100 A journalist for the Guardian who covered events in Bosnia-Herzegovina in the period between 1992 and 1995.
101 *Mladić* TC (Chapter 1, n 105) § 4504.
102 *Mladić* TC (Chapter 1, n 105) § 4505.

with the Geneva Conventions, with 'detainees being given three meals a day and not being forced to work at the front lines'.[103] In its factual and legal findings, the Trial Chamber found, *inter alia*, that detainees had to perform various manual labour tasks at the front lines and that 'non-Serb detainees lost significant amounts of weight, with a number of them dying of malnutrition due to a shortage of food'.[104] It also found that 'camp guards frequently and severely beat detainees'.[105] Accounting for Mladić's position at the time these statements were made, and in view of his contemporaneous assurance that detainees were under his control, the Trial Chamber thus found his statements to have been deliberately misleading.

136. This view was supported by evidence that prior to visits of two Serb detention camps by ICRC representatives and foreign journalists, the camps were arranged and prepared with measures taken to make the conditions inside the camps appear satisfactory.[106] Pursuant to an order made by Mladić, standards inside detention camps were improved prior to these visits 'while detainees were hidden or otherwise made inaccessible for interviews by journalists or inspection by the ICRC'.[107] In both cases, the Trial Chamber determined that 'Mladić's words and actions were deliberately misleading as they were made in an attempt to portray the camp conditions in a more favourable light than what was actually the case'.[108]

137. When considering whether Mladić's conduct in this respect – alongside other acts and omissions – amounted to a significant contribution to the Overarching JCE, the Trial Chamber found that Mladić's communicative acts were so instrumental to the commission of the crimes that without them the crimes would not have been committed as they were. The judgement referenced how Mladić 'introduced and maintained a controlled and centralised system of spreading propaganda related to Bosnian Croats and Bosnian Muslims [and] made deliberately misleading statements to members of the media and international community in relation to crimes committed on the ground.'[109]

138. The Trial Chamber found that Mladić 'significantly contributed to achieving the objective of permanently removing the Bosnian Muslims and Bosnian Croats from Bosnian Serb-claimed territory in Bosnia-Herzegovina

103 *Mladić* TC (Chapter 1, n 105) § 4511.
104 *Mladić* TC (Chapter 1, n 105) § 4511.
105 *Mladić* TC (Chapter 1, n 105) § 4512.
106 *Mladić* TC (Chapter 1, n 105) §§ 4507, 4508.
107 *Mladić* TC (Chapter 1, n 105) § 4512.
108 *Mladić* TC (Chapter 1, n 105) § 4512.
109 *Mladić* TC (Chapter 1, n 105) § 4498.

through persecution, extermination, murder, inhumane acts (forcible transfer), and deportation'.[110] Further, it noted that given Mladić's knowledge of the aforementioned crimes, he 'should have ensured that crimes carried out by the VRS and subordinated forces were investigated, and perpetrators properly punished'.[111] Finding that Mladić deliberately misled the international community on the conditions inside detention camps and attempted to portray them in a more favourable light, however, the Trial Chamber determined Mladić's actions were 'indicative of his overall stance towards investigating war crimes and crimes against humanity' and found that he had not taken appropriate steps to investigate or punish perpetrators of crimes.[112]

139. In addition, in evaluating the evidence that Mladić gave orders to the VRS to respect the Geneva Conventions and sometimes appeared to pursue peaceful solutions to the conflict, the Trial Chamber considered this alongside his other actions and statements.[113] Both in light of events contemporaneously unfolding on the ground and the deliberate spread of false information by Mladić, the Trial Chamber found that his 'orders to respect the Geneva Conventions, his statements to UNPROFOR personnel, and his involvement in peace negotiations were not indicative of his true state of mind'.[114] Based on all of the foregoing, the Trial Chamber found that the Accused intended to permanently remove Bosnian Muslims and Bosnian Croats from Bosnian Serb claimed territory through the commission of the crimes of deportation, inhumane acts (forcible transfer), murder, extermination, and persecution. The Trial Chamber is satisfied that the Accused shared the intent to achieve the common objective of the Overarching JCE through the commission of these crimes and found that the Accused held this intent by 12 May 1992 at the latest.

140. At the *Mladić* Appeals Trial before the IRMCT, the subject of Mladić's spread of disinformation again came up through several points raised. The Appeals Chamber recalled that in evaluating Mladić's statements, conduct, and knowledge of crimes to determine his *mens rea* in relation to the Overarching JCE, the Trial Chamber relied on considerable evidence to concluded that he shared a criminal intent. This included 'Mladić's repeated use of derogatory terms to refer to Bosnian Muslims and Bosnian Croats, his recalling of historical crimes allegedly committed against Bosnian Serbs, his expressions of commitment to an ethnically homogeneous Republika Srpska, and his

110 *Mladić* TC (Chapter 1, n 105) § 4612.
111 *Mladić* TC (Chapter 1, n 105) § 4546.
112 *Mladić* TC (Chapter 1, n 105) § 4546.
113 *Mladić* TC (Chapter 1, n 105) § 4646.
114 *Mladić* TC (Chapter 1, n 105) § 4687.

provision of misinformation while knowing about the commission of crimes in the Municipalities'.[115] As such, the Appeals Chamber found that Mladić did not demonstrate that the Trial Chamber had erred in their use of contextual evidence in assessing his *mens rea*.

141. Although Mladić objected on appeal that the Trial Chamber failed to give sufficient weight to the genuine warnings in his orders for VRS soldiers to respect the Geneva Conventions,[116] the Appeals Chamber found that these were 'not indicative of his true state of mind' as these were directly contradicted by 'what happened on the ground', his provision of disinformation, and 'his other contemporaneous statements'.[117]

142. Similarly, in assessing Mladić's orders to observe ceasefire agreements the Appeals Chamber concluded that his involvement in peace negotiations was not indicative of his state of mind given that 'these actions and statements, sometimes providing misinformation, [were] inconsistent with [his] other conduct and [were] directly contradicted by his other contemporaneous statements'.[118]

143. The *Mladić* case thus illustrates several ways in which the dissemination of disinformation has been linked to the commission of atrocity crimes by an international tribunal. These include: an evocation of historical narratives of Serb victimisation intended to mobilise Serbs through engendering fear and hatred; false information provided to the public about VRS intentions and activities which facilitated the commission of ongoing crimes; and the deliberately misleading of the international community about conditions inside detention centres which was seen as indicative of Mladić's overall stance towards investigating and punishing the commission of war crimes and crimes against humanity. The Appeals Chamber found that Mladić's use of disinformation during the commission of international crimes constituted compelling evidence in determining his knowledge and intent and it found that his disinformation – alongside other acts and omissions – represented one of the ways in which he contributed to the overarching JCE.

3.5 Case Vignette: Karadžić at the ICTY

144. The trial of Radovan Karadžić, president of the Republika Srpska, at the ICTY included substantial evidence of disinformation, ranging from distracting

115 *Mladić* AC (Chapter 1, n 96) § 257.
116 *Mladić* AC (Chapter 1, n 96) § 259.
117 *Mladić* AC (Chapter 1, n 96) § 259.
118 *Mladić* AC (Chapter 1, n 96) § 260.

the international community to covering up specific crimes. Notably, there was considerable discussion of misleading public statements made by Karadžić which differed from the reality unfolding on the ground, despite his awareness of this fact, implying intentional duplicity in his dealings with the international community. The Trial Chamber considered whether and how such misleading statements played a role in creating an environment of impunity and as such encouraged and contributed to the ongoing implementation of the common plan to commit war crimes and crimes against humanity, as argued by the Prosecution.[119] The speech acts of the defendant were not simply window dressing: 'The Chamber finds that these speeches and statements went beyond mere rhetoric and formed a core element in the policies and plans developed by the Accused and the Bosnian Serb leadership'.[120]

145. The Trial Chamber found that Karadžić 'frequently minimised the extent of criminal activity in the Municipalities when speaking to the international media, negotiators, and to the public'.[121] This included claims that 'civilians were leaving of their own free will and signed voluntary declarations that they wanted to leave' as well as statements made to international media that 'Bosnian Serb authorities did not participate in ethnic cleansing but were trying to stop this practice'.[122] He further asserted that 'acts of ethnic cleansing were being perpetrated by groups or individuals who did not originate from the Bosnian Serb authorities'.[123] At the time when Karadžić made such statements, for example when he stated in an April 1992 media interview that there were very few Bosnian Muslims and Bosnian Croats living in Serb areas who had been expelled, he was aware of the reality that 'that thousands of non-Serb civilians had been expelled from Bosnian Serb controlled territory'.[124]

146. The Trial Chamber referenced multiple ways in which Karadžić spread false and misleading information about the conduct of Serb forces, including regarding the commission of crimes. It was noted that, throughout the conflict, Karadžić 'boasted to the Bosnian Serb Assembly about the proper conduct of the Serb forces, when he knew it was not true'.[125] Further, Karadžić stated in 1993 that 'the UN could present evidence of war crimes' but the Republika Srpska not only 'could never have committed crime' but also should 'investigate

119 *Prosecutor v Radovan Karadžić* (Judgement) IT-95-5/18-T (24 March 2016) TC § 3471.
120 *Karadžić* TC (Chapter 2, n 119) § 3486–3487.
121 *Karadžić* TC (Chapter 2, n 119) § 3376.
122 *Karadžić* TC (Chapter 2, n 119) § 3376.
123 *Karadžić* TC (Chapter 2, n 119) § 3376.
124 *Karadžić* TC (Chapter 2, n 119) § 3376.
125 *Karadžić* TC (Chapter 2, n 119) § 3378.

and prosecute matters itself'.[126] In addition, Karadžić 'shifted blame for the crimes away from Bosnian Serbs' by claiming that 'sources were biased', 'his interlocutors had been misinformed', 'crimes and terrible things had happened historically and were continuing to happen to Bosnian Serbs as well', 'everybody was guilty' and 'the international community was responsible'.[127]

147. Karadžić made false statements about the presence of civilians who had been forcibly removed from their homes inside of Serb detention centres. He stated on multiple separate occasions – including in speeches and television interviews – that 'we don't have people in prisons that have been removed from their own homes. Those people have been captured in the battle field', that 'there were no civilians in our prisons', and that none of the institutions who had inspected the prisons found any civilians or women and indeed that a large number of detainees had been released.[128] Prior to making these statements, he had already been informed that a large number of civilians were being detained in detention centres with inadequate conditions both through a report commissioned by the Bosnian Serb government and by representatives of the international community and the media.[129] As such, with respect to the crime of unlawful detention, the Trial Chamber found that Karadžić was aware of the unlawful detention of non-Serb civilians from conflict areas.[130] It was noted, however, that he only made efforts to close them down when the international media began reporting on inhumane conditions in detention centres while continuing to 'present to the public that the allegations made by the Guardian that the Bosnian Serbs had organised concentrations camps or were holding civilian prisoners were completely false'.[131]

148. Regarding his knowledge of ethnic cleansing, the Trial Chamber found that despite learning of such instances throughout the conflict, Karadžić only issued instructions to the Bosnian Serb forces to prevent the forced resettlement of civilians on 19 August 1992, many months after the start of the conflict. Karadžić followed these instructions with a letter to the international community on 25 August 1992 in which he explained said instructions stating that 'the civilian population must be allowed to move freely out of a war zone if that was the desire of the civilian population' and blaming the international community for failing to understand the 'deep-rooted antagonism' and ethnic

126 *Karadžić* TC (Chapter 2, n 119) § 3379.
127 *Karadžić* TC (Chapter 2, n 119) § 3377.
128 *Karadžić* TC (Chapter 2, n 119) § 3377–8.
129 *Karadžić* TC (Chapter 2, n 119) § 3377.
130 *Karadžić* TC (Chapter 2, n 119) § 3378.
131 *Karadžić* TC (Chapter 2, n 119) § 3498.

hatred driving displacement.[132] Supported by the fact that Karadžić 'continued to share the intent to forcibly remove non-Serbs from the Municipalities until 30 November 1995' the Trial Chamber found the statements made by Karadžić in his letter both to be an attempt at covering up and minimising the forced nature of removals and indicative of the fact that the orders he issued 'to prevent such forced removals were not intended.'[133]

149. The Trial Chamber also discussed, in relation to sniping and shelling campaigns, the submission by the Prosecution that although Karadžić 'would sometimes acknowledge responsibility for the sniping and shelling of civilians, generally he denied and deflected international criticism, including through false assurances, false denials or excuses, blaming others, cavalier brush-offs, using *tu quoque*, or threatening to do worse' for the purpose of advancing such campaigns.[134] Numerous witness, including primarily representatives of the international community, provided evidence 'that they or their colleagues protested and complained to [Karadžić] about the sniping and shelling of the civilian population of Sarajevo'.[135] For example, one witness testified that when asked directly why the Bosnian Serbs continued to shell Sarajevo, Karadžić responded that they were not stating 'it's the Muslims. We are not attacking, just protecting our homes in and around Sarajevo.'[136] Despite his apparent full awareness of the large-scale commission of crimes and numerous appeals by the international community to cease action against civilian populations, Karadžić did not issue an order to stop the attack.[137]

150. The Trial Chamber also recalled that 'throughout the relevant period, [Karadžić] actively monitored the international media coverage of the events in Srebrenica', embarking together with Mladić 'on an effort to disseminate false information about the fate of Bosnian Muslim males as well as the conditions under which the remainder of the Bosnian Muslim population was transferred to Potočari.'[138] Karadžić denied international organisations the ability to access Srebrenica and other areas such as Bratunac and Zvornik, where crimes were being committed. As such, the Trial Chamber took the view that due to Karadžić's 'contemporaneous knowledge of the ongoing killing operation, the only reasonable inference is that by disseminating false information,

132 *Karadžić* TC (Chapter 2, n 119) § 3496.
133 *Karadžić* TC (Chapter 2, n 119) § 3496.
134 *Karadžić* TC (Chapter 2, n 119) § 4811.
135 *Karadžić* TC (Chapter 2, n 119) § 4813.
136 *Karadžić* TC (Chapter 2, n 119) § 4813.
137 *Karadžić* TC (Chapter 2, n 119) § 4813.
138 *Karadžić* TC (Chapter 2, n 119) § 5812.

[Karadžić] intended to shield the true actions of the Bosnian Serb Forces from international attention and intervention.'[139]

151. Further, when confronted with allegations of rape and expulsions of Bosnian Muslim civilians, Bosnian Serb leaders did not deny that such crimes were taking place but rather attempted to suggest that they were being committed 'by Bosnian Muslims who were masquerading as Serb soldiers'.[140] Karadžić himself stated on 8 January 1993 that these allegations were 'terrible lies' as 'no soldier of ours would be allowed to rape a woman in the presence of another soldier'.[141] On 9 January 1994 he went so far as to make a statement to the Bosnian Serb Assembly that 'contrary to the lies and slander of international propaganda-mongers, the Serbian army maintained the knightly character and military honour worthy of it'.[142]

152. These examples highlight the numerous ways in which Karadžić disseminated disinformation to shift blame for crimes committed by Bosnian Serbs,[143] 'minimised what he knew about criminal activity committed by Serb forces,' to mislead the public and the international community,[144] and to deny or cover up or crimes (a form of aiding and abetting). The Trial Chamber found that there was a 'clear disjuncture between the manner in which [Karadžić] defended the actions of the Bosnian Serbs in international settings and press conferences and the reality on the ground, of which he was fully aware'.[145] It therefore concluded that:

> At the same time that he was learning about crimes committed against non-Serbs and not taking sufficient steps to prevent or punish them, [Karadžić] was providing misleading information to international observers on the ground and the media. By his denials that Bosnian Serbs were committing crimes in the Municipalities and his disingenuous portrayal of the reality on the ground, of which he was in fact fully aware, [Karadžić] created an environment in which Bosnian Serbs could continue to commit the crimes through which the common purpose of the Overarching JCE was implemented.[146]

139 *Karadžić* TC (Chapter 2, n 119) § 5812.
140 *Karadžić* TC (Chapter 2, n 119) § 3377.
141 *Karadžić* TC (Chapter 2, n 119) § 3378.
142 *Karadžić* TC (Chapter 2, n 119) § 3380.
143 *Karadžić* TC (Chapter 2, n 119) § 3377.
144 *Karadžić* TC (Chapter 2, n 119) § 3381.
145 *Karadžić* TC (Chapter 2, n 119) § 3503.
146 *Karadžić* TC (Chapter 2, n 119) § 3504.

153. Karadžić made several allegations of errors in the Trial Chambers assessment of his statements being misleading and his duplicity in dealing with the international community, which were dismissed by the Appeals Chamber at the IRMCT.[147]

154. In *Karadžić*, the dissemination of disinformation concealed crimes for the purpose of preventing international intervention and to facilitate impunity that permitted the commission of further crimes in the absence of investigation or punishment. Several examples illustrate how Karadžić spread false narratives to deflect criticism of the VRS and deny its crimes against Bosnian Muslim civilians by shifting blame and deliberately misleading the public and international community.

3.6 Case Vignette: Mbarushimana Confirmation of Charges at the ICC (Pre-Trial Chamber)

155. In the ICC prosecution case against Callixte Mbarushimana, Executive Secretary of the *Forces Démocratiques pour la Libération du Rwanda* (FDLR), the Pre-Trial Chamber declined to confirm the charges against the defendant. Although the *Mbarushimana* case never went to trial after the majority of the Pre-Trial Chamber judges decided that the contextual elements that provide the basis for crimes against humanity charges were not established, it nevertheless provides an important reference point on how the dissemination of disinformation might contribute to the commission of mass atrocity crimes – particularly through their concealment. To analyse the alleged disinformation disseminated by Mbarushimana, the judges developed a two-pronged test which has important potential for a wider use in future cases concerning disinformation. As such, although the charges against Mbarushimana himself were not confirmed, this case requires further discussion due to the relevance and potential application of the two-pronged test to other cases. Furthermore, the dissenting opinion of Presiding Judge Monageng and emerging debate on the potential role played by disinformation in the case provide important insight into the contentious and unresolved nature of this issue in international criminal law.

156. The *Mbarushimana* confirmation of charges decision considered attacks on civilian populations by the FDLR, resulting in the commission of several war crimes and crimes against humanity, with the purpose of creating a 'humanitarian catastrophe' in the Kivu provinces of Democratic Republic of

147 *Karadžić* AC (Chapter 1, n 89) § 413.

Congo (DRC) in January 2009. It also considered whether these were attacks which were carried out pursuant to a strategy to 'draw the world's attention to the FDLR's political demand'.[148] The Prosecution alleged that Mbarushimana, in his capacity as Executive Secretary of the FDLR, was responsible for implementing the strategy of publicly and systematically denying any FDLR responsibility for attacks on civilians, whilst having full knowledge of FDLR attacks perpetrated against the civilian population.[149] The Prosecution thus submitted that Mbarushimana 'was the "linchpin" of the overall FDLR strategy, namely because of his ability to transform the FDLR's crimes on the ground into political capital' and by 'exploiting the international attention on the FDRL and its political agenda triggered by the attacks'.[150] Thus, the prosecution case theory is essentially one of aiding and abetting war crimes by a media and publicity coordinator for an armed organisation committing mass atrocities.

157. The media campaign run by Mbarushimana, which he himself referred to as 'a war of information'[151] consisted of two aspects running parallel to one another. On the one hand, the Prosecution alleged that Mbarushimana was – in his civilian capacity and based in Europe – 'publicly, immediately, repeatedly, vehemently and falsely deny[ing] the FDLR's direct involvement in the crimes and attribute them to other national or local armed groups'.[152] At the same time, he was making statements to use as leverage in order to pressure the international community by insisting that 'the waves of civilian casualties would inevitably continue unabated until the States negotiated with and resolved the FDLR's legitimate grievances'.[153] In the Prosecution's view, the combination of 'Mbarushimana's extortive negotiation demands and his public denials were his criminal and voluntary contribution to the FDLR's common purpose'.[154]

148 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) §§ 6–7.

149 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 8.

150 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) §§ 7–8.

151 *Prosecutor v Callixte Mbarushimana* (Prosecution's Written Submissions on the Confirmation of Charges) ICC-01/04-01/10-448-Red (06 October 2011) § 5, referring to DRC-OTP-2038–2153 [EVD-PT-OTP-00782], 25 January 2009 email from Mbarushimana to Murwanashyaka and others referring to the 'guerre de communication / information contre l'ennemi qui nous a attaqué' and proposing the publication of daily 'situation reports' to complement the press releases. See also the 35 press releases listed in Slide 13, shown at the hearing, from which the Prosecution submitted it is evident that Mbarushimana believed that he was waging such a 'media war' by denying the FDLR's involvement in crimes and shifting the overall responsibility for the attacks on civilians onto the Congolese and Rwandan governments.

152 *Mbarushimana* Prosecution's Written Submissions (Chapter 2, n 151) § 5.

153 *Mbarushimana* Prosecution's Written Submissions (Chapter 2, n 151) § 5.

154 *Mbarushimana* Prosecution's Written Submissions (Chapter 2, n 151) § 5.

158. Accordingly, the Prosecution submitted that Mbarushimana should be held criminally responsible under common purpose liability through the application of Article 25(3)(d) of the Rome Statute having knowingly and intentionally contributed 'in any other way' to the commission of war crimes and crimes against humanity.[155] The Prosecution alleged that Mbarushimana had contributed to the common plan 'by agreeing [...] to conduct an international media campaign as part of the common plan'.[156]

159. The Majority of the Chamber, with Judge Monageng dissenting, eventually declined to confirm the charges, finding that there was insufficient evidence to establish that Mbarushimana, 'denied crimes committed by the FDLR with knowledge of them and in furtherance of a policy of the organisation'.[157] They therefore determined that, within the meaning of Article 25(3)(d) of the Rome Statute, he could not have provided a significant contribution to the commission of FDLR crimes 'through his radio communications and press releases'.[158]

160. In arriving at this conclusion, the Majority employed a two-pronged test to evaluate and determine when the concealment or denial of crimes is not neutral. This test consisted of the following two elements: (i) the suspect has awareness of the crimes committed and knows that he is denying the truth; and (ii) this denial of the truth is done in furtherance of a policy or common purpose.[159]

161. Regarding the first element relating to the defendant's awareness of crimes committed and knowledge that truth is being denied, the Majority noted that 'evidence is contradictory as to whether Mr Mbarushimana, or even the FDLR political leadership, was aware of the crimes committed on the ground'.[160] Not everything occurring on the ground 'was reported to the leadership, especially when it comes to crimes against civilians', there were contradictions in witness statements and an overall lack of clarity about reporting and communication channels within the FDLR leadership.[161]

162. When it came to the second element relating to a determination whether the aim behind denying the truth was to further the common purpose, the Majority took the view that 'the evidence is not sufficient to demonstrate that

155 *Mbarushimana* Prosecution's Written Submissions (Chapter 2, n 151) § 5. As charged by the Prosecution under Counts 1–13 – see also § 108 which lists out the counts.
156 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 293.
157 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 315.
158 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 315.
159 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 312.
160 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 313.
161 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 313.

the Suspect denied crimes in furtherance of a policy of the organisation'.[162] It was noted by the Majority that 'the arguments raised [...] to support allegations that [Mbarushimana] was aware of the commission of crimes and that denial of these crimes was a policy of the organisation are based on assumptions about [his] knowledge of the alleged crimes at the time when he distributed his radio communications and press releases' but that substantial grounds to determine Mbarushimana had such knowledge was not established by the Prosecution.[163]

163. A key issue in the case was the extent to which Mbarushimana's concealment and denial of the attacks on civilians by FDLR troops had an impact on the audience (therefore constituting a perlocution) who received Mbarushimana's press statements. The Majority was persuaded that because the press releases prepared and issued by Mbarushimana were primarily intended for the international community 'they did not have an impact on troops on the ground'.[164] Although the Prosecution presented evidence that some press releases were passed to the troops, 'depending on their content and on their likely effect on the soldiers' morale', the Majority ultimately concluded that the press releases were not illocutionary; that is, intended to encourage soldiers to commit crimes against the civilian population.[165]

164. In her dissenting opinion, presiding Judge Monageng expressed her disagreement with the majority as to the existence of the contextual elements of crimes against humanity and thus provided her view on the charges against Mbarushimana presented by the Prosecution.[166] This included an analysis of Mbarushimana's contribution to the common purpose under Article 25(3)(d) liability which specifically considered his press releases aimed at concealment of crimes and press releases intended to encourage FDLR troops.[167] Regarding the two-pronged test employed by the Majority to determine when denying crimes is not 'neutral', Judge Monageng disagreed with the Majority's findings coming to the conclusion that the test had in fact been met because Mbarushimana was aware of the falsity of his statements and had made them pursuant to an organisational policy.[168]

165. Judge Monageng disagreed with the view taken by the Majority on the basis of witness statements that 'Mbarushimana had no influence on the

162 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 314.
163 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 314.
164 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 328.
165 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 328.
166 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 27.
167 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent §§ 65–112.
168 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 76.

soldiers in the field'.[169] Notably, in Judge Monageng's opinion 'the fact that these witnesses state that Callixte Mbarushimana did not have power, control or authority over FDLR commanders and soldiers does not, in and of itself, preclude a finding that [he] made a significant contribution to the commission of crimes by FDLR soldiers'.[170] Crucially, Monageng's dissent notes that under Article 25(3)(d) of the Rome Statute – contribution in any other way – such power or authority is not required and a contribution to the crime need only be 'significant' but not 'essential'.[171]

166. Again, in the *Mbarushimana* Pre-Trial Chamber decision there was disagreement between the Majority and presiding Judge Monageng on the extent of the defendant's contribution to the crimes. While the Majority did not consider that there was sufficient evidence indicating substantial grounds to continue the case against Mbarushimana, Judge Monageng stated in her dissent that in the context of an armed conflict in the Democratic Republic of the Congo in which there were mass civilian casualties, Mbarushimana's actions show that he 'was using the international media campaign to conceal past crimes and encourage future ones'.[172] In this light, the *Mbarushimana* decision exhibits a difference of opinion between the judges about the relevant stage of the speech act; while Judge Monageng seemed to be focusing on the illocutionary dimensions of the speech acts of the accused (i.e. what he sought to encourage), the majority focused on the perlocutionary aspects (or, what was achieved).

167. It is also worth noting that in the *Mbarushimana* case, the Defence made reference to the acquittal of Hans Fritzsche from the IMT, seemingly suggesting that 'it stands for the proposition that a propagandist merely denying crimes cannot be held criminally responsible under international law'.[173] This cross-fertilisation of propaganda jurisprudence from two temporally and geographically distinct situations, is illustrative both in relation to the continuing role of propaganda driving atrocities and for the ongoing principled dispute over denial of crimes as a basis for criminal liability. In Judge Monageng's dissent, the Defence's reliance on the *Fritzsche* trial was misplaced. Given the controversy of the *Fritzsche* acquittal even at the time, she contended that it should not be understood as standing for a general 'proposition that, under international criminal law, no one should be held responsible

169 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 83.
170 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 83.
171 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent §§ 84, 96.
172 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 112.
173 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 67.

for denying crimes via propaganda'.[174] Judge Monageng proceeded to highlight some of the key differences that emerge between the *Fritzsche* case and the *Mbarushimana* case. Notably, Mbarushimana 'had authority over shaping the FDLR's media campaign, discretion as to what was said and [...] knowledge of the criminal activity of his organisation'.[175] It is also worth noting that although Fritzsche was acquitted at the IMT, he was later convicted and incarcerated by West Germany.

168. Ultimately, the *Fritzsche* Judgement was premised on the lack of sufficient evidence and lack of mode of liability (such as complicity) within the IMT framework that could adequately capture a contribution in the form of disinformation. This should, however, not be interpreted as a definite conclusion that 'international criminal law precludes responsibility when propagandists deny crime'.[176] There are other cases in which the misleading denial of crimes has been a factor contributing to convictions, including *Mladić* and *Karadžić*, as discussed herein.

169. Although charges were not confirmed in *Mbarushimana*, there is nevertheless useful judicial reasoning in the decision and dissent that can inform our analysis of disinformation in international criminal law. In particular, the two-pronged test employed by the judges in the *Mbarushimana* decision to evaluate the neutrality of statements denying a crime provides a basis for further development, and the potential basis for evaluating the contribution of various types of disinformation to the commission of international crimes. Furthermore, the discussion in *Mbarushimana* highlights the importance of analysing perlocutions (or consequences) and assessing the contributions of disinformation to increasing the likelihood of a criminal offence being committed. Ultimately, the debates in *Mbarushimana* are indicative of the disagreements that exist over how to weigh the contribution of disinformation to the commission of international crimes. This highlights the complexity of establishing a causal link (beyond a reasonable doubt) between a disinformation campaign aimed at the international community or media pronouncements occurring in one context and atrocity crimes unfolding in another.

3.7 Case Vignette: Ruto and Sang Confirmation of Charges at the ICC

170. The ICC case of Kenyan President William Samoei Ruto and Kalenjin-language Kass FM radio broadcaster Joshua Arap Sang, concerned with crimes

174 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 110.
175 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 110.
176 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) Judge Monageng Dissent § 110.

against humanity (of murder, deportation or forcible transfer of population and persecution)[177] allegedly committed in the context of post-election violence in Kenya in 2007 and 2008, also contains a disinformation element. The charges against Ruto and Sang were ultimately vacated in 2016, without prejudice to their future prosecution, due to concerns over systematic tampering with witnesses and concerns about the admission of evidence. Although the case was terminated, reviewing the Confirmation of Charges and other key trial documents, in particular discussions relating to the role of Sang, nevertheless raises important points regarding the potential role disinformation campaigns can play in inciting to violence and contributing to crimes against humanity.

171. With respect to his criminal responsibility for crimes against humanity, the Prosecutor charged Sang under the mode of liability contained in Article 25(3)(d) of the Rome Statute – contribution 'in any other way.'[178] This was the same mode of liability in the ICC Prosecution's charges against Mbarushimana. As the Chamber considered evidence regarding Sang's contribution to the commission of crimes against humanity they evaluated, among other aspects, his 'broadcasting of false news regarding alleged murders of Kalenjin people in order to inflame the atmosphere in the days preceding the elections'.[179] The Chamber took the view that 'there [were] substantial grounds to believe' Sang had intentionally contributed to the aforementioned crimes against humanity.[180] This finding was supported by the testimony of multiple witnesses.[181]

172. Specifically regarding the spread of disinformation, the Chamber considered the testimony of Witness 8 which stated that 'between 3 November 2007 and 27 December 2007, Mr. Sang broadcasted false news regarding murders or more general offences allegedly committed against members of the Kalenjin community in order to instil fear within the Kalenjins and prepare them to fight the enemy communities'.[182] In addition, according to Witness 1, Sang had asserted through Kass FM in the days leading up to the election 'that Kikuyus

177 *Prosecutor v William Samoei Ruto, Henry Kiprono Kosgey and Joshua Arap Sang* (Decision on the Confirmation of Charges Pursuant to Article 61(7)(a) and (b) of the Rome Statute) ICC-01/09-01/11 (23 January 2012) § 350.

178 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 351.

179 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 355. These other aspects included placing his show at the disposal of the organisation, advertising the meetings of the organisation, fanning the violence through the spread of hate messages explicitly revealing desire to expel the Kikuyus, and broadcasting instructions during the attacks in order to direct the physical perpetrators to the areas designated as targets.

180 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 357.

181 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 357.

182 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 357.

would rig the election'.[183] This was considered in conjunction with other inciting statements allegedly broadcast by Sang immediately before the eruption of violence.[184]

173. Sang made similar statements about the stolen election as soon as the electoral results were announced, as reported by Witness 2 who stated that Sang said that 'the elections had been stolen and our rights denied' and that people 'should get their weapons from where they were [sic] kept and, if necessary, to use any arm at their disposal to evict the Kikuyus'.[185] This was corroborated by testimony from Witness 4, who affirmed that, following the announcement of the newly elected President, Sang told his listeners that 'they should resist the Kibaki administration because Kibaki had stolen the votes' and that 'Kikuyu in Eldoret and Kapsabet had been attacked and rightly so because the votes had been stolen'.[186] Further, Witness 1 heard Sang on the air for Kass FM asking Kalenjin listeners 'why aren't you going out to stop the Kikuyus?'.[187]

174. The Defence presented evidence to rebut the allegation that Sang contributed to the commission of crimes against humanity by collecting witness statements which denied that he had 'broadcasted any inciting messages' stating instead that he had 'only disseminated peaceful appeals'.[188] The Chamber, however, observed 'that a large number of the witness statements presented appear to be drafted in a systematic manner' using 'exactly the same wording or employ[ing] sophisticated legal terminology, which is not common for non-lawyers'.[189] The Chamber continued, 'these statements were mostly collected during the same day' and mostly did not provide precise temporal reference but rather 'confined themselves to generic expressions' when referring to the 2007/2008 period of post-election violence.[190] Additionally, the

183 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 357.

184 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 358. For example, according to the testimony of Witnesses 2 and 8, Sang said that 'if Kibaki wins, we will carry out our work' (Statement of Witness 2, KEN-OTP-0053-0256 at 0269) and 'we will give the instructions' (Statement of Witness 8, KEN-OTP-0052-0880 at 0883). Witness 2 then elaborated on the word 'work', explaining that this term was used in the three months preceding the election instead of explicitly using to the word 'kill' (Statement of Witness 2, KEN-OTP-0055-0035 at 0043). Witness 2 also further clarified that 'to carry out the work' meant to make sure that Kikuyus 'have been evicted [...][and] have been killed' (Statement of Witness 2, KEN-OTP-0055-0035 at 0043).

185 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 359.

186 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 359.

187 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 359.

188 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 363.

189 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 361.

190 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 361.

Chamber was not convinced that this evidence rebutted Sang's contribution due to 'the possibility that beside these peaceful messages, the instructions and speeches reported by Witnesses 1, 2, 4, 6 and 8 could have also been broadcasted, despite of the existence of the ban'.[191] This conclusion was supported by 'detailed explanations regarding the content, nature and timing of [Sang's] broadcasting during the period immediately preceding the election and in the course of the commission of the crimes' provided by these witnesses.[192] Despite this affirmation in the confirmation decision, Sang would go on to be ultimately cleared after his no case to answer submission was upheld by the Trial Chamber.

4 Analysis – Disinformation in International Criminal Law

175. The case law set out above identifies the following features as relevant to a legal determination of criminal liability in relation to the dissemination of disinformation: the significance of the statement maker, with more authoritative and influential figures able to generally create greater harm through the spread of disinformation; the intentionality of the dissemination of false information, such as to cover up crimes and mislead the media; and the amplification of the disinformation.

176. Additional points in this respect include that disinformation can occur before a crime to encourage crimes (e.g. false flag operations to justify a war of aggression), during crimes (such as the disinformation during operations designed to deter NATO from intervening in the forced displacement of civilians in the *Gvero* case), or in the aftermath of crimes (such as in the *Mladić* and *Karadžić* examples where false statements where used both during and immediately after the crimes to mislead the press and prevent international actors from taking measures to intervene in and/or investigate the events). In each of these situations the Prosecution must prove, beyond a reasonable doubt, that disinformation contributed to the commission of atrocity crimes. In many such cases, there may not be a linear and proximate sequence of causation between the disinformation and the commission of a crime, and the expectation of a causal-relationship with the occurrence of crimes is unrealistic, as disinformation is seldom, if ever, the *sine qua non* cause of an international crime. Nonetheless, it may be jointly sufficient with other causal factors to result in a crime, and it may be sufficient on its own to amount to a culpable contribution

191 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 363.
192 *Ruto and Sang* Confirmation of Charges (Chapter 2, n 177) § 363.

to a crime. In such circumstances, the most appropriate modes of liability are aiding and abetting and contributing to an organised criminal group, as well as inducing or soliciting, depending on the specific circumstances of the case. Proving liability on the basis of sharing disinformation will generally require an evidentiary package indicating the state of mind in relation to the crime of the person spreading the disinformation and their willingness to assist or encourage the perpetration or cover-up of crimes.

177. Following the *Bemba* and *al-Hassan* Judgements at the ICC, the *actus reus* contribution threshold for aiding and abetting is low but the *mens rea* standard is heightened. Determining whether disinformation itself is part of the commission of a crime such as persecution would require meeting the contextual elements of crimes against humanity and the necessary elements of the underlying act. To charge liability for aiding and abetting, at a minimum, the person spreading the disinformation would have to be aware that the information was untrue and that it was distributed to advance the underlying crimes. However, if pursuing aiding and abetting is infeasible, other modes of liability such as ordering, soliciting, or inducing may nonetheless be appropriate. However, disinformation in and of itself, in the absence of accompanying mass crimes, does not qualify as a prompting another to commit a crime for the purposes of Article 25(3)(b) of the Rome Statute – as is discussed in further detail in the following section.

178. Disinformation can also be analysed in relation to direct and public incitement to genocide. When considering whether disinformation could constitute incitement, the directness requirement may pose difficulties for the Prosecution. Often disinformation sustains genocide by supporting the inciting message and an environment conducive for genocide and it may exacerbate direct and public incitement to genocide.[193] But to qualify as direct and public incitement to genocide, a disinformation campaign would have to convey a message that was clearly understood by the listeners as a direct call for genocide of one of the four protected groups. In assessing directness, it is necessary to take account of potential temporal and the geographic links, as well as the overarching context in which the disinformation campaign occurs. This analysis of the felicity conditions for uptake of locutions or illocutions that are systematic falsehoods should also factor in whether disinformation was disseminated in isolation or whether it was spread as part of a larger narrative, for example immediately following multiple direct calls to violence at a rally. As overarching considerations, it is important analytically to distinguish between disinformation as evidence for an established crime, which is a

193 See discussion in the section Case Vignette: Karadžić at the ICTY.

separate concern to disinformation potentially constituting a distinct crime in and of itself, which is a normative question.

179. Although the approach to disinformation in international criminal law remains a contentious issue, there are several important points of discussion that emerge from the overview of cases presented here. The starting point for any analysis is whether the disinformation campaign and intentionally distorted message met the threshold and made a significant or even substantial contribution to the likelihood of crimes occurring. In other words, the intentionally untrue message was designed to provoke ground-level perpetrators to carry out mass atrocity crimes or to deter external entities who could prevent the commission of such crimes from intervening. This possibility has been evaluated in key international trials in different ways and though the international case law is indicative of certain guiding principles, it is as yet unsettled.

180. Disinformation campaigns may evoke alleged historical crimes committed by the other side, often with exaggeration, used to create a narrative of threat and fear. In this way, 'disinformation is used most straightforwardly as a propaganda tool to condition and encourage the future commission of atrocity crimes'.[194] Disinformation campaigns in the form of cover-ups of ongoing crimes may also contribute to the commission of further crimes or at least create an environment in which such crimes can continue unabated when they are intended to prevent intervention by the international community or through the creation of an environment of impunity in which ongoing crimes are neither investigated nor punished. This also raises the issue of the temporal link between a criminal act committed and the spread of disinformation which seeks to deny said act after the fact, which is discussed further in Chapter 2, including in the context of domestic genocide denial laws.

4.1 Disinformation as a Contribution to a Crime (Modes of Liability)

181. The overview of cases with a disinformation element presented above demonstrates the different ways in which disinformation may contribute to the commission of atrocity crimes and the modes of liability through which the dissemination of disinformation may have contributed to the commission of a crime. Some of the challenges of accurately capturing the contribution of disinformation to international crimes emerged as early as the trial of Fritzsche at the IMT-Nuremburg, which did not explicitly contain accessorial liability within its legal mandate.

194 Holvoet (Chapter 2, n 3) 225.

182. The mode of liability of aiding and abetting is found in Article 25(3)(c) of the Rome Statute and is also now consistently included across the Statutes of international criminal tribunals.[195] Aiding and abetting has a lower causation requirement than other modes of liability, making it perhaps the most adept manner of conceptualising the contribution of 'non-military propagandists and demagogic politicians to a collective criminal enterprise'.[196] In addition, the 'in any other way' form of liability contained in Article 25(3)(d) of the Rome Statute could also potentially be employed in cases involving a disinformation element brought before the ICC.

4.2 *Actus reus* (the Physical Element)

183. The mode of liability of aiding and abetting consists of both acts or omissions (where there is a duty to act) which provide assistance, encouragement, or moral support to the perpetration of crimes. These acts or omissions must have a substantial effect on or make a substantial contribution to the commission of an offence. The threshold of 'substantial effect' for liability under aiding and abetting is higher than that of 'significant contribution or effect' through the joint criminal enterprise, which is discussed further below. Of particular relevance to speech crimes is that aiding and abetting, which is generally understood as the 'lending of practical assistance in the commission of a crime' also encompasses the provision of 'encouragement or moral support'.[197] Aiding and abetting can be rendered at any point in relation to the offence – before, after, or during its commission – and a cause and effect relationship is not a necessary requirement.[198]

4.3 *Mens rea* (the Mental Element)

184. At the ICTY and ICTR, the *mens rea* standard for aiding and abetting was knowledge, not purpose, in relation to the commission of the crime: 'In the

195 Rome Statute (Key Guidelines, n 2) art 25(3)(c); UN Security Council, *Statute of the International Criminal Tribunal for the Former Yugoslavia* (adopted 25 May 1993) UN Doc S/RES/827 (1993) art 7(1); and UN Security Council, *Statute of the International Criminal Tribunal for Rwanda* (adopted 8 November 1994) UN Doc S/RES/955 (1994) art 6(1).

196 Wilson and Gillett (Key Guidelines, n 1) § 178.

197 *Prosecutor v Nikola Šainović, Nebojša Pavković, Vladimir Lazarević, Sreten Lukić* (Judgement) IT-05-87-A (23 January 2014) § 1626, § 1649; *Prosecutor v Anto Furundžija* (Judgement) IT-95-17/1-T (10 December 1998) §§ 233–235; *Prosecutor v Radoslav Brđanin* (Judgement) IT-99–36-T (01 September 2004) § 271. See Wilson and Gillett (Key Guidelines, n 1) § 180.

198 See *Prosecutor v Tihomir Blaškić* (Judgement) IT-95-14-A (29 July 2004) § 48, Wilson and Gillett (Key Guidelines, n 1) §§ 183–185, and Wilson (Chapter 2, n 34) 286–287.

case of aiding and abetting, the requisite mental element is knowledge that the acts performed by the aider and abettor assist the commission of a specific crime by the principal.'[199] In international criminal tribunal Judgements, 'knowledge' for complicity implies only that the accomplice was aware that a crime would be committed in the ordinary course of events by the direct perpetrators; that is, that a crime was a possible and foreseeable consequence of his or her conduct.[200] In the hierarchy of mental states, awareness (*dolus eventualis*) of a risk of a harm eventuating is lower than knowledge (at least as interpreted), and furthermore, international tribunal judgements have determined that the accomplice's awareness of the exact nature of the crime may be only partial: it is not necessary that the aider or abettor know the precise crime that was intended or which was actually committed. If he is aware that one of a number of crimes will probably be committed, and one of those crimes is in fact committed, he has intended to assist or facilitate the commission of that crime and is guilty as an aider or abettor.[201] The *mens rea* element of aiding and abetting stands in stark contrast to the 'common plan, design, or purpose'[202] of joint criminal enterprise – the other more elevated form of accessorial liability.

185. The ICC Statute dramatically elevated the *mens rea* element for aiding and abetting compared to the interpretation of this mode of liability by the ad hoc tribunals. Article 25(3)(c) of the Rome Statute holds a person criminally responsible for a crime if that person, 'for the purpose of facilitating the commission of such a crime, aids, abets or otherwise assists in its commission or its attempted commission, including providing the means for its commission.'[203] By adding the language 'for the purpose,' the Statute introduces a heightened mental element of purpose that verges on conscious desire or motive and in practice, markedly raises the *mens rea* requirement. Comparatively, the

199 *Prosecutor v Dusko Tadić* (Judgement) IT-94-1-A (15 July 1999) § 229. The same language is repeated in *Prosecutor v Mitar Vasiljević* (Judgement) IT-98-32-A (25 February 2004) § 102(ii); *Blaškić* Charges (n 321) § 49; and *Prosecutor v Miroslav Kvočka, Milojica Kos, Mlado Radić, Zoran Žigić, Dragoljub Prcać* (Judgement) IT-98-3-/1-T (02 November 2001) § 253.

200 *Kvočka* TC (Chapter 2, n 199) § 255.

201 *Prosecutor v Blagoje Simić* (Judgement) IT-95-9-A (28 November 2006) § 86, citing the authority of *Blaškić* AC (Chapter 2, n 198) § 50, which opined, '[i]t is not necessary that the aider and abettor has knowledge of the precise crime that was intended' or that was actually committed, as long as he was aware that one of a number of crimes would probably be committed, including the one actually perpetrated. This language is repeated in *Kvočka* TC (Chapter 2, n 199) § 255.

202 In establishing the mental elements for JCE, *Tadić* AC (Chapter 2, n 199) § 227 refers to: 'The existence of a common plan, design or purpose to commit a crime.'

203 Rome Statute (Key Guidelines, n 2) art 25(3)(c).

'purpose' of the ICC Statute is sharply divergent from the knowledge standard used for complicity in most criminal jurisdictions.[204]

186. Aiding and abetting has yet to be fully adjudicated at the ICC, so there is still a degree of uncertainty as to its interpretation. However, the ICC's Pre-Trial Chamber in *Mbarushimana* stated that 'the jurisprudence of the ad hoc tribunals does not require the aider and abettor to share the intent of the perpetrator to commit the crime, whereas under article 25(3)(c) of the Statute, the aider and abettor must act with the purpose of facilitating the commission of that crime.'[205] The purpose standard was a central justification for the Pre-Trial Chamber's decision in *Mbarushimana* to refuse to confirm the charges against the defendant. The refusal is notable given that Mbarushimana was the spokesman and propagandist for the *Forces Démocratiques pour la Libération du Rwanda* (FDLR),[206] who issued communiqués stating that 'the price of a country is blood not water'[207] at a time when the FDLR was systematically murdering civilians and creating a humanitarian catastrophe in the eastern Democratic Republic of Congo. This highlights the Pre-Trial Chamber's strict interpretation of Article 25(3)(c), while also suggesting that the judges exercised particular caution where the Accused's contribution was made through expression rather than the physical provision of weapons or other material. That strict interpretation was also followed in the *Al-Hassan* Trial Judgement, with the Chamber noting that the wording of Article 25(3)(c) 'introduces a "higher subjective mental elements" than that provided for in Article 30 of the Statute and means that the accessory must have lent his or her assistance with the aim of facilitating the crime.'[208]

187. If applied strictly, the ICC's purpose test for complicity is likely to hinder the charging and prosecuting of professional wartime propagandists such as Mbarushimana under a theory of aiding and abetting. The limiting function of this strict interpretation in the digital era will have far-reaching ramifications. The advance of technology has opened up new avenues to contribute remotely to atrocity crimes. Terrorist organisations such as the Islamic State have already shown the power of digital tools on social media platforms for recruitment, influencing, and disruption of society's orderly functioning. Groups like these

204 However, it is consistent with the American Law Institute, *Model Penal Code: Official Draft and Explanatory Notes* (ALI 1962) s2.06(4).

205 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) § 281.

206 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) §§ 274, 291.

207 *Mbarushimana* Confirmation of Charges (Chapter 2, n 26) §§ 321.

208 *Prosecutor v Al-Hassan* (Judgement) ICC-01/12-01/18 (26 June 2024) § 1227.

are increasingly incorporating AI into their terroristic endeavours.[209] More broadly, individuals contributing to movements that embrace the use of violence against civilians can amplify their personal impact through digital and AI tools, such as agentic AI.

188. If the potential liability of such individuals and groups for aiding and abetting is restricted to those who can be proved to have shared the purpose of facilitating the crime, many of those who substantially and knowingly contribute to crimes will potentially escape liability by arguing that their purpose was instead monetary, political, or religious, for example. In this light, the PJI Guidelines endorse adherence to the long-standing *mens rea* standard of knowledge that is the settled law of aiding and abetting at international criminal tribunals and in the preponderance of national criminal jurisdictions.[210] While a reasonable *mens rea* element is surely needed, 'purpose' is overly restrictive of the scope of potential liability, particularly in the digital era. There is widespread support in the secondary legal literature that international criminal law should retain knowledge as the intent requirement of aiding and abetting.[211] Stated plainly, the accomplice need not share the specific purpose to commit an international crime, but he or she must have knowledge of the material perpetrators' intention to commit a crime and nonetheless decide to proceed with providing the contribution to the foreseeable crime.[212]

5 Disinformation in the Form of Genocide Denial and Its Treatment in Rwanda and Germany

5.1 Genocide Denial in Rwanda

189. Aside from disinformation designed to further ongoing current atrocities, disinformation is sometimes deployed in revisionist ways to alter accounts of even the most notorious historical crimes. For example, Rwanda and Germany were the settings for two genocidal campaigns of the twentieth century. Both

209 See Arianna Vadaschi and Chiara Graziani, *Artificial Intelligence in Counter-Terrorism: Human Rights Challenges and Institutional Implications* (Edward Elgar 2025).

210 Elies Van Sliedregt, *Individual Criminal Responsibility in International Law* (Oxford University Press 2012) 129.

211 See, e.g. Miles Jackson, *Complicity in International Law* (Oxford, Oxford University Press 2015) 80; Van Sliedregt (Chapter 2, n 210) 129.

212 An approach adopted in *Krstić* AC (Chapter 1, n 10) § 142, that Schabas (Executive Summary, n 3) 351 summarises but criticises in favour of a lesser *mens rea* standard.

States have taken active measures to suppress denial of those genocides, including through the imposition of criminal liability for genocide denial. These examples provide guidance for the regulation of disinformation in countries with established historical records of experiencing the worst crimes of concern to the international community.

190. Regarding Rwanda, the right to freedom of expression is enshrined in Rwandan domestic law through Article 38 of the 2003 Constitution. Article 38 imposes an important limitation by stating that 'the right to freedom of expression and freedom of access to information shall not prejudice public order, good morals, the protection of the youth and children, the right of every citizen to honour and dignity and protection of personal and family privacy.'[213] Genocide denial is addressed in Rwanda through the framework of law No °59/2018 (a revision of law No °18/2008). Under this law, relating to the punishment of the crime of genocide ideology, anyone found to be minimising, denying, or otherwise justifying genocide is liable to a prison sentence of five to nine years and a substantial monetary fine.[214]

191. The broader human rights framework applicable to Rwanda includes the African Charter on Human and People's Rights (ACHPR), which provides for freedom of expression in Article 9.[215] Although Article 9 does not directly contain possible limitations or restrictions on this right, Article 27(2) contains a general limitations clause requiring that all rights and freedoms 'shall be exercised with due regard to the rights of others, collective security, morality and common interest'.[216] A further potential means for defining restrictions on the right to freedom of expression as enshrined in Article 9 can be found in Article 28 of the ACHPR,[217] which states that 'every individual shall have the duty to respect and consider his fellow beings without discrimination, and to maintain relations aimed at promoting, safeguarding and reinforcing mutual respect and tolerance'.[218]

192. In practice in Rwanda, the authorities have applied this law in a manner that has been found to be over-broad or ambiguous and thereby in violation of human rights protections for freedom of expression. These concerns have been addressed by the African Court on Human and Peoples' Rights (AfCHPR) in its

213 Constitution of the Republic of Rwanda 2003 (as amended 2015) art 38.

214 Law No 59/2018 of 22 August 2018 on the Crime of Genocide Ideology and Related Crimes (Rwanda).

215 African Charter on Human and Peoples' Rights (adopted 27 June 1981, entered into force 21 October 1986) (1982) 21 ILM 58 (African Charter) art 9.

216 African Charter (Chapter 2, n 215) art 27(2).

217 Wilson and Gillett (Key Guidelines, n 1) § 50.

218 African Charter (Chapter 2, n 215) art 28.

2017 case *Ingabire Victoire Umuhoza v. The Republic of Rwanda*. In *Umuhoza v. Rwanda*, the African Court asserted that assessments of whether restrictions on freedom of expression are necessary and proportional 'cannot be done in a vacuum and due consideration should be given to particular contexts in which the impugned expressions were made'.[219] In its judgement, the Court acknowledged that given Rwanda's history 'it is entirely legitimate for the state to have introduced laws on the "minimisation", "propagation", or "negation" of genocide'.[220] In *Umuhoza v. Rwanda*, however, the African Court determined that the statements made by the applicant neither minimised nor denied the genocide, and as such her conviction and sentencing was 'not necessary in a democratic society' and was in violation of Article 9(2) of the ACHPR and Article 19 of the ICCPR.[221] The Court also noted that even if it were to accept the necessity of the restrictions imposed on the applicant's statements, the punishment imposed by the State was not proportionate and less restrictive measures could have been adopted for the same objective.[222]

193. Given the role of incendiary hate speech and media in the 1994 Rwanda genocide, there is in principle a legitimate basis to enact measures limiting freedom of expression pertaining to genocidal ideology and with the aim of preserving social cohesion. Such measures have been found to serve a legitimate purpose and have been considered necessary for the protection of public order, as in the African Court's judgement in *Umuhoza v. Rwanda*. However, the jurisprudence of human rights bodies has warned about the risks of suppressing disfavoured speech in practice and has highlighted that 'the laws in question should not be applied at any cost to the rights and freedoms of individuals or in a manner which disregards international human rights standards'.[223]

5.2 Memory Laws in Germany

194. Turning to Germany, the broader regional context includes the European Convention on Human Rights (ECHR), which enshrines freedom of expression in Article 10.[224] Subject to restrictions under Article 10(2) ECHR, any restrictions imposed on freedom of expression must not only be proportionate

219 *Ingabire Victoire Umuhoza v. The Republic of Rwanda* App no 003/2014 (AfCHPR, 24 November 2017) § 144.

220 *Umuhoza v Rwanda* (Chapter 2, n 219) § 147.

221 *Umuhoza v Rwanda* (Chapter 2, n 219) §§ 162–163.

222 *Umuhoza v Rwanda* (Chapter 2, n 219) § 162.

223 *Umuhoza v Rwanda* (Chapter 2, n 219) § 148.

224 *Convention for the Protection of Human Rights and Fundamental Freedoms* (European Convention on Human Rights, as amended) CETS No 005 (ECHR) art 10.

to the pursuit of a legitimate aim, but also must (a) be prescribed by law, (b) have a legitimate aim, and (c) be considered necessary in a democratic society because of a 'pressing social need'.[225] Through Article 17 ECHR, which prohibits the abuse of rights enshrined in the Convention, the case law of the European Court of Human Rights (ECtHR) has addressed the issue of Holocaust denial and incitement of hatred against an ethnic group and has held on multiple occasions that imposing a prison sentence for such crimes can be lawful.[226]

195. Article 5 of the German Constitution (Grundgesetz – German Basic Law) codifies the right to freedom of expression, subject to certain limitations[227] – including through Section 130 of the German Criminal Code (Strafgesetzbuch) which addresses incitement of masses (Volksverhetzung).[228] This encompasses 'publicly or in a meeting approve[ing] of, den[ying] or downplay[ing] an act committed under the rule of National Socialism [which qualifies as genocide] in a manner suited to causing a disturbance of the public peace.'[229] Genocide denial can be penalised with a maximum of five years imprisonment. The German Criminal Code also provides that '[w]hoever publicly or in a meeting disturbs the public peace in a manner which violates the dignity of the victims by approving of, glorifying or justifying the National Socialist tyranny and arbitrary rule incurs a penalty of imprisonment for a term not exceeding three years or a fine'.[230] The German parliament (*Bundestag*) expanded the offence of incitement in October of 2022,[231] to allow up to three years imprisonment

225 See, e.g. *Jersild v Denmark* App no 15890/89 (ECtHR, 23 September 1994); *Lehideux and Isorni v France* App no 24662/94 (ECtHR 23 September 1998); *Otegi Mondragon v Spain* App no 2034/07 (ECtHR 15 March 2011); *Faruk Temel v Turkey* App no 16853/05 (ECtHR 1 February 2011).

226 See, e.g. *Garaudy v France* App no 65831/01 (ECtHR 24 June 2003), where the ECtHR held that a six-month prison sentence for holocaust denial was compatible with art 10 ECHR (Chapter 2, n 224), as interpreted in light of art 17 ECHR. See similarly *Molnar v Romania* App no 16637/06 (ECtHR 2012). See further ECtHR, 'Guide on Article 17 of the European Convention on Human Rights – Prohibition of abuse of rights' (updated 28 February 2026).

227 Grundgesetz für die Bundesrepublik Deutschland (GG – German Basic Law) 1949 (Germany) art 5(2).

228 Strafgesetzbuch (StGB – German Criminal Code) 1871 (Germany) s130(1): A person disturbs public peace if they incite 'hatred against a national, racial, religious group or a group defined by their ethnic origin, against sections of the population or individuals on account of their belonging to one of the aforementioned groups or sections of the population or calls for violent or arbitrary measures against them'.

229 StGB (Chapter 2, n 228) s130(3).

230 StGB (Chapter 2, n 228) s130(4).

231 StGB (Chapter 2, n 228) s130(5). The background to the rapid amendment of the law is an infringement proceeding brought against the Federal Republic of Germany by the EU

for individuals who – in a manner which incited hatred or violence against a person or group in order to disturb the public peace – approve of, deny or grossly trivialise war crimes, crimes against humanity, or genocide against a group or an individual because of their affiliation with a group.[232]

196. Germany is not the only country in Europe which punishes the denial of the Holocaust, incitement to genocide and in some cases, also incitement to war crimes and crimes against humanity – although it does impose some of the highest maximum prison sentences.[233] The State's interpretation of when and how a certain severity threshold for speech crimes is met may be influenced and dependent on its social and historical context, including the prior occurrence of a genocide. Some such laws require that a clear link can be established to incitement to hatred of violence in order to justify incarceration.[234] Where domestic laws are broader, there is a risk that States will misuse such provisions to punish any speech considered offensive. This concern was noted by the AfCHPR in *Umuhoza v. Rwanda*, when it affirmed that 'putting severe restrictions, such as criminal punishments, on the rights of individuals merely on the basis of contexts would create an atmosphere where citizens cannot freely enjoy basic rights and freedoms, including the right to freedom of expression.'[235]

197. Also of relevance is the ECtHR Judgement in *Salov v Ukraine 65518/01*, in which Salov spread untrue claims about the death of a presidential candidate.[236] The proportionality test for Article 10 ECHR failed as the ECtHR found that domestic courts had not proved intent and the impact of Salov's statements were limited, given that only a small number of copies were distributed. The penalty imposed on Salov for his false statements was thus determined to be too severe and consequently, a violation of Article 10 ECHR. This case suggests both that statements whose truthfulness can be called into

---

Commission in December 2021. The Commission had criticised Germany for inadequate implementation of Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, particularly with regard to public denial and gross trivialisation.

232 StGB (Chapter 2, n 228) s130(5).

233 Germany, Romania, Portugal, Italy, and Austria impose the highest maximum prison sentences – with five (Germany, Romania and Portugal), six (Italy), and ten (Austria) years. See European Parliamentary Research Service, 'Holocaust denial in criminal law – Legal frameworks in selected EU Member States' (2021).

234 As is the case in Germany.

235 *Umuhoza v Rwanda* (Chapter 2, n 219) § 159.

236 *Salov v Ukraine* (Chapter 2, n 5).

question are nevertheless protected under Article 10, and that the dissemination of information – even where it is strongly suspected to be false – is not prohibited.[237] However, genocide denial in countries that have experienced such events may be repressed through criminal sanction as long as it is not applied in an overly broad or arbitrary manner. The ECtHR has consistently found that Holocaust denial is not protected by the ECHR, for example in the cases *Pastörs v. Germany 55224/14* and *Gauraudy v. France 65831/01*.

237 Urška Umek, 'Council of Europe's responses to disinformation' (*Council of Europe* 14 May 2021) <https://www.osce.org/sites/default/files/f/documents/2/3/487525.pdf> accessed 26 June 2026.

CHAPTER 3

Social Media and Speech Crimes

198. Disinformation and hate speech can be spread through traditional forms of media or broadcasting platforms – including radio, television, newspapers, and similar outlets. The growth of social media in recent years has meant that disinformation and hate speech can proliferate at scale and across the globe instantaneously. On the one hand, this provides a tool for violent and incendiary messages to be easily and quickly communicated to a large audience of potential perpetrators of material crimes.[1] On the other, social media provides new avenues and possibilities for holding to account perpetrators, instigators, and those aiding and abetting or otherwise responsible for atrocity crimes if evidence can be gathered, collected, and stored as they are taking place.[2] While this may in some ways provide access to previously inaccessible situations of conflict from a distance and allow for communication with victims from afar, it also comes with its own challenges and limitations.[3]

199. This section will first provide a brief overview of the pre-digital case of *Nahimana et al.*, which centred on incendiary speech that incited the Rwandan genocide, along with several more recent case studies in which speech crimes allegedly spread through the use of social media have played a seemingly important role in contributing to the commission of mass atrocity crimes. This is followed by a discussion of emerging accountability gaps, the risk of an environment of impunity around social media, and potential avenues towards accountability for social media companies, including whether the use of automated algorithmic models and recommender systems might have implications for criminal or civil liability. Given its necessity for accountability, this section is followed by a more in-depth presentation on the gathering and submitting of evidence of speech crimes, which implies further reflection on evidence in the digital era.

1 For more on this subject refer also to Sarah Zarmsky, 'Is International Criminal Law Ready to Accommodate Online Harm? Challenges and Opportunities' (2024) 22:1 Journal of International Criminal Justice 169, 172–176.

2 For more on this subject of the dual role played by social media not only as it pertains to speech crimes but international criminal law more generally, see Rebecca J Hamilton, 'Social Media Platforms in International Criminal Investigations' (2020) 52:1 Case Western Reserve Journal of International Law 213.

3 Hamilton (Chapter 3, n 2) 218.

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-005
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

1 A Pre-Digital Era Case

1.1 Case Vignette: Nahimana et al. (the Media Case) at the ICTR

200. Because the ICTR jurisprudence on topics linked to hate speech and social media was examined in the Hartford Guidelines,[4] it is not repeated in extensive detail here. Nonetheless, the key findings in *Nahimana et al.* usefully identify the contributing role which the forerunners of social media – namely traditional print and broadcast media, designed for mass audiences – can play to atrocity crimes.

201. The *Nahimana et al.* case, which focused on the role played by media (and especially broadcast radio) during the Rwandan genocide, involved three accused: Ferdinand Nahimana, an academic who helped to establish the radio station *Radio Télévision Libre des Mille Collines* (RTLM) used to spread hate speech and propaganda; Jean-Bosco Barayagwiza, a lawyer who also helped to establish the RTLM and who was a founding member of the *Coalition pour la défense de la République* (CDR); and Hassan Ngeze, a journalist and founding member of the CDR party who started the Kangura newspaper, of which he later became Editor-in-Chief. The three accused utilised both radio (the RTLM) and print media (the *Kangura* newspaper) to disseminate harmful and even genocidal messages. Their messages were patently outside of the scope of public interest and therefore fell squarely into the type of prohibited speech that can attract criminal responsibility.

202. Some of the messages spread by the accused directly called for the commission of atrocities, while other messages used indirect language to contribute to and encourage the genocidal targeting of Tutsis. Both RTLM and *Kangura* repeatedly conveyed messaging which framed Tutsi women as 'seductive agents of the enemy', thereby increasing the likelihood that they would be targeted and killed or become victims of sexual and gender-based violence.[5] To understand the intent of the accused in spreading these messages, the Trial Chamber considered the purpose of publicly transmitting such material through the RTLM radio station and the Kangura newspaper, and whether it was of a 'bona fide nature' (meaning here for the purpose of historical research, the dissemination of news and information, or for promoting public accountability of government authorities).[6]

4 Wilson and Gillett (Key Guidelines, n 1) §§ 201 et seq.
5 *Nahimana et. al* TC (Introduction, n 2) § 1079.
6 Wilson and Gillett (Key Guidelines, n 1) § 86.

203. Due to the immediate and active nature of radio transmissions, the Trial Chamber found messages spread through the RTLM to be especially dangerous and harmful, as they created a heightened sense of fear, danger, and urgency which in turn created a need to take action among listeners.[7] A broadcast from 4 June 1994 stood out to the Trial Chamber as particularly illustrative of the ways in which the RTLM engaged in direct and public incitement of genocide. One noteworthy example involved the radio presenter Habimana telling his listeners to eliminate the *Inkotanyi*, who they would recognise by their height and physical appearance. Statements such as '[j]ust look at his small nose then break it' vividly symbolised the intent to destroy the Tutsi ethnic group, according to the Trial Chamber.[8]

204. The Chamber found that the killing of Tutsis was promoted by the CDR, which Barayagwiza assisted to found. This was evidenced by the chanting of 'tubatsembatsembe' or 'let's exterminate them', by Barayagwiza himself, other CDR members and *Impuzamugambi* (Rwandan militia of young Hutus) in his presence at meetings and demonstrations.[9] The killing of Tutsi civilians was also promoted by the CDR through the publication of communiqués and other writings that called for the extermination of the enemy and defined the enemy as the Tutsi civilians.[10]

205. The Trial Chamber recognised that many writings published in the *Kangura* newspaper combined ethnic hatred and fear-mongering with a call to violence against the Tutsi population, who were characterised as the enemy or enemy accomplices. As the founder, owner, and chief editor of *Kangura*, Ngeze directly controlled the content of its publications and was responsible for them, a fact that he acknowledged. Editorials published in *Kangura* included statements like '[…] blood will really flow. All the Tutsis and the cowardly Hutus will be exterminated'.[11] An article titled 'Appeal to the Conscience of the Hutu' contained 'The Ten Commandments', which included such statements as 'if we fail to achieve our goal, we will use violence' and 'The Hutu should not take pity on the Tutsi'. One witness explained that they understood these

---

7 *Prosecutor v Ferdinand Nahimana, Jean-Bosco Barayagwiza, and Hassan Ngeze* (Judgement and Sentence) ICTR-99-52-T (03 December 2003) § 1031 (*Nahimana et al.*).

8 *Nahimana et al.* TC (Introduction, n 2) § 1032.

9 *Nahimana et al.* TC (Introduction, n 2) §§ 336, 964.

10 *Nahimana et al.* TC (Introduction, n 2) § 964 'The genocidal intent in the activities of the CDR was expressed through the phrase "tubatsembasembe" or "let's exterminate them", a slogan chanted repeatedly at CDR rallies and demonstrations. At a policy level, CDR communiques called on the Hutu population to "neutralize by all means possible" the enemy, defined to be the Tutsi ethnic group.'

11 *Nahimana et al.* TC (Introduction, n 2) § 225.

commandments as preparation for the killings to come. The article claimed that the enemy was 'waiting to decimate us' and called on Hutus to 'wake up' and 'take all necessary measures to deter the enemy from launching a fresh attack'.[12] The Trial Chamber found that the text of this article was 'an unequivocal call to the Hutu to take action against the Tutsi'.[13]

206. These pre-digital incidents of incitement to genocide and other inflammatory statements, along with the earlier examples of Streicher and the ICTY cases, show that the use of mass media channels to encourage atrocities has a long history. Yet, the advent of the digital era increases the reach and occurrence of such incendiary campaigns, as is demonstrated in the following vignettes. Linguistically, the adjudication of speech crimes can be demanding; particularly in relation to areas with less access to digital technology and/or languages that are not widely known, for which challenges arise 'including the absence of trained language professionals for many of the target languages, the lack of existing lexical items to denote international legal concepts, the languages' frequent lack of written tradition, and low rates of literacy in victim communities'.[14]

2 Incitement to Genocide and Other Atrocity Crimes on Social Media

207. At first, many observers applauded the democratising potential of social media. In 2011 and 2012, pro-democracy movements in Egypt, Syria, Tunisia, and Russia organised mass protests against authoritarian regimes on Facebook.[15] In Latin America, civil society activists quickly mobilised on social media against government corruption and human rights violations.[16] However, governments soon took to social media and employed digital technologies to consolidate

12 *Nahimana et al.* TC (Introduction, n 2) § 153.

13 *Nahimana et. al* TC (Introduction, n 2) § 153.

14 Swigart (Chapter 1 , n 30) 275.

15 See Ron Deibert, 'Authoritarianism Goes Global: Cyberspace Under Siege' (2015) 26:3 Journal of Democracy 64, 65; Zeynep Tüfekçi, 'Social Movements and Governments in the Digital Age: Evaluating a Complex Landscape' (2014) 68 Journal of International Affairs 1; Zeynep Tüfekçi, 'How Social Media Took Us From Tahrir Square to Donald Trump' (*MIT Technology Review* 14 August 2018) <https://www.technologyreview.com/2018/08/14/240325/how-social-media-took-us-from-tahrir-square-to-donald-trump/> accessed 24 June 2026.

16 Leopoldo Fergusson and Carlos Molina, 'Facebook Causes Protests' (*CEDE* 2019).

state power, and in some cases, coordinate international crimes and violate human rights.[17]

208. Since the advent of social media, there has been an upsurge in extreme nationalist and nativist political ideology[18] in mainstream politics, globally increasing the risk of heightened violence against minority groups.[19] Data suggest a connection between this rise in rhetoric to increases in hate crimes, though empirical studies have reported mixed results in this respect.[20] A 2019 report by the then-UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression emphasised the empirical connection between hate speech and incitement to violence in contemplating social media companies' hate speech policies,[21] and this nexus has been confirmed in the empirical social science literature on hate speech.[22]

2.1 Case Vignette: Rohingya in Myanmar

209. The targeting of the Rohingya in Myanmar stands out as a significant example of the harmful use of social media by a government and has been considered part of a potentially genocidal campaign by the United Nations Independent International Fact-Finding Mission on Myanmar (IIMM). According

17 See Tüfekçi (Chapter 3, n 15) 23; Deibert (Chapter 3, n 15) 65; Sam Gregory, 'Cameras Everywhere Revisited: How Digital Technologies and Social Media Aid and Inhibit Human Rights Documentation and Advocacy' (2019) 11 Journal of Human Rights Practice 373.

18 See Ralph Schroeder, *Social Theory After the Internet: Media, Technology, and Globalisation* (UCL Press 2018) 60 asserting that digital media were a necessary precondition for the rise of right-wing and nationalist movements in China, India, Sweden, and the US.

19 Arcila Calderón, C., Sánchez Holgado, P., Gómez, J. *et al.* 'From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and LGBT communities' Humanit Soc Sci Commun 11, 1369 (2024).

20 Cf. May Bulman, 'Brexit Vote Sees Highest Spike in Religious and Racial Hate Crimes Ever Recorded' (*The Independent* 07 July 2017) <https://www.independent.co.uk/news/uk/home-news/racist-hate-crimes-surge-to-record-high-after-brexit-vote-new-figures-reveal-a7829551.html> accessed 16 June 2026 with Pablo Madriaza et al. 'Exposure to hate in online and traditional media: A systematic review and meta-analysis of the impact of this exposure on individuals and communities' (2025) Campbell systematic reviews vol. 21,1 e70018.

21 David Kaye, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (09 October 2019) UN Doc A/74/486 § 8.

22 See David Yanagizawa-Drott, 'Propaganda and Conflict: Evidence from the Rwandan Genocide' (2014) 129 Quarterly Journal of Economics 1947, 1989; Karsten Müller and Carlo Schwarz, 'From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment' (2023) 15(3) American Economic Journal: Applied Economics 270; Ayal Feinberg, Regina Branton and Valerie Martinez-Ebers, 'The Trump Effect: How 2016 Campaign Rallies Explain Spikes in Hate' (2022) 55(2) PS: Political Science & Politics 257.

to a report on the situation in Myanmar issued by the Office of the Prosecutor of the International Criminal Court, anti-Muslim rhetoric and threats of violence against Rohingya were posted on Facebook by high-ranking members of Myanmar's military and Buddhist nationalists beginning in 2012.[23] In two waves of violence in 2016 and 2017, military and non-Rohingya civilians violently attacked Muslim Rohingya villages in Rakhine State.[24] The most conservative estimate is that the Myanmar armed forces (*Tatmadaw*) killed 10,000 Rohingya civilians,[25] but some observers put the number higher.[26] As of September 2018, there had been 6,097 documented incidences of gender-based violence, including sexual violence,[27] and approximately 780,000 Rohingya had been forcibly displaced to neighbouring Bangladesh.[28]

210. It is generally accepted that posts on Facebook incited specific acts of violence against the Rohingya and contributed to a general climate that made the atrocities against them possible. A UN-sponsored fact-finding mission documented over 150 public social media accounts, pages, and groups on Facebook that 'regularly spread messages amounting to hate speech against Muslims in general or Rohingya in particular.'[29] The report noted a number of these accounts were 'particularly influential' due to 'the number of followers (all over 10,000, but some over 1 million), the high levels of engagement of the followers with the posts (commenting and sharing), and the frequency of new posts (often daily, if not hourly).'[30] The Myanmar government deployed online narratives evoking several different themes about the supposed threat posed by the Rohingya Muslims, including descriptions of them as 'an existential threat to the country,' 'a threat to Burmese racial purity,' and 'a threat to

23 Office of the Prosecutor (OTP), Request for authorisation of an investigation pursuant to article 15, 'Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar' (4 July 2019) ICC-01/19 § 59.

24 OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) §§ 5, 67–68.

25 OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) § 90.

26 OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) § 89.

27 OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) § 94.

28 OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) § 69.

29 'Report of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar' (17 September 2018) UN Doc A/HRC/39/CRP.2 § 1310 (Myanmar Report, Detailed Findings).

30 HRC IIMM Report, Detailed Findings (Chapter 3, n 29).

Buddhist religious sanctity.'[31] According to the UN mission, '[s]uch narratives latch onto long-standing anti-Muslim prejudices and stereotypes; they are designed to stoke fear.'[32]

211. There is ample evidence to suggest that these campaigns were influential in inciting specific attacks. For example, while a widespread and systematic attack on the Rohingya civilian population was underway, young soldiers in the Myanmar military posted derogatory and inciting messages on Facebook containing the same narratives deployed by propaganda accounts. On August 26, 2017, a soldier wrote that he could not wait to be deployed to Rakhine State because the 'Muslim dogs' represented a threat to the citizens, and the next day he posted, 'on the battlefield, whoever is quick will get to eat you, Muslim dogs.'[33] On August 27, 2017, a police officer involved in forcibly removing and killing Rohingya civilians posted that he '[h]a[s] been wanting to kill these *kalar* for so long. Only got to kill them just now.'[34]

212. In its assessment of the situation, the Office of the Prosecutor of the International Criminal Court noted: 'A lieutenant of the 33rd Light Infantry Division (LID) deployed to Rakhine in early August 2017 posted a "smirking emoji" on Facebook, along with the commentary, "[i]f they're Bengali [...] they'll be killed."'[35] In a report released in 2018, Reuters documented more than 1000 posts, comments, images, and videos on Facebook attacking the Rohingya or Muslims generally in extremely derogatory terms and threatening them with violence.[36] All the posts were in the Burmese language. Further, the UN mission affirmed that in addition to direct incitement, anti-Rohingya speech on social media 'contributed to and exacerbated a climate in which hate speech thrives and in which individuals and groups may be more receptive to calls of incitement to violence.'[37]

213. The impact of derogatory messages was amplified by the lack of alternative channels for information and low levels of digital literacy in Myanmar. The dependence of the population on Facebook for internet access is so great that, '[t]he word "Facebook" has become synonymous with the internet itself in

31 HRC IIMM Report, Detailed Findings (Chapter 3, n 29) §§ 1315–1317.

32 HRC IIMM Report, Detailed Findings (Chapter 3, n 29) § 1314.

33 OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) §§ 71, 176.

34 '*Kalar*' is an anti-Muslim racial epithet. OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) § 177.

35 OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) §§ 71, 176–177.

36 OTP, Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Chapter 3, n 23) § 59.

37 HRC IIMM Report, Detailed Findings (Chapter 3, n 29) § 1327.

Myanmar.'[38] The Report of the Independent International Fact-Finding Mission on Myanmar concluded, '[t]he role of social media is significant. Facebook has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the Internet.'[39] The IIMM placed heavy emphasis on derogatory statements against Rohingya in determining that there is a reasonable basis to find genocidal intent behind the atrocities committed against them. The use of Facebook by Myanmar's military to incite and pursue atrocities against the Rohingya is widely acknowledged, as Facebook/Meta publicly accepted responsibility for its central role in these acts, which the Independent International Fact-Finding Mission on Myanmar has found, applying a reasonable grounds test, could be characterised as genocidal.[40]

2.2 Case Vignette: Russia and Ukraine

214. From late 2013 to early 2014, a period of internal political unrest in Ukraine sparked the 'Euromaidan' mass movement – signalling Ukraine's decisive change in political course away from Russia and towards Europe.[41] This culminated in the removal of then President Yanukovych under European Union (EU) mediation which was met with further protests in the Eastern regions of the country – particularly in Crimea.[42] In late February 2014, armed and (mostly) uniformed local and Russian military personnel alongside members of local militia groups took *de facto* control of the Crimean Peninsula.[43] A sham referendum was held on 16 March 2014, and on 18 March 2014 the Russian Federation formally announced the incorporation of Crimea into

38 Megha Rajagopalan and others, 'How Facebook Failed the Rohingya in Myanmar' (*BuzzFeed News* 27 August 2018) <https://www.buzzfeednews.com/article/meghara/facebook-myanmar-rohingya-genocide> (accessed 11 June 2026).

39 HRC IIMM Report (Chapter 2, n 9) § 74.

40 HRC IIMM Report, Detailed Findings (Chapter 3, n 29) §§ 1388–1441. See also Paul Mozur, 'A Genocide Incited on Facebook, with Posts from Myanmar's Military' (*New York Times* 15 October 2018) <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html> accessed 11 June 2026; Alexandra Stevenson, 'Facebook Admits It Was Used to Incite Violence in Myanmar' (*New York Times* 6 November 2018) <https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html> accessed 24 June 2026.

41 International Federation for Human Rights (FIDH), Kharkiv Human Rights Protection Group, and Center for Civil Liberties, 'Communication under Article 15 of the Rome Statute: Situation in Ukraine – Hate Speech as the Crime Against Humanity of Persecution' (June 2024) § 39.

42 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) § 40.

43 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) § 41.

Russian territory, which it has continued to occupy ever since.[44] Commenting on the annexation, Putin and the Russian media have cited the need to protect the rights and lives of ethnic Russians and Russian speaking residents of Crimea from 'nationalists, neo-Nazis, Russophobes and anti-Semites'.[45] However, Russia has stubbornly resisted an international peace-keeping force in the Donbas, despite the fact that such a force would have the very task of protecting civilians from the harms alleged by the Russian State.

215. After the full-scale invasion by the Russian Federation of Ukraine on 24 February of 2022, political observers,[46] international non-governmental organisations[47] and a UN Commission of Inquiry[48] reported on the significant role played by an intricate web of disinformation and propaganda, including the widespread use of hate speech against Ukrainians in traditional media such as state-owned TV and on social media and other platforms.

216. A recent communication submitted under Article 15 of the Rome Statute made by the International Federation for Human Rights (FIDH) together with the non-governmental organisations Kharkiv Human Rights Protection Group and Center for Civil Liberties alleges that 'virulent hate speech that denigrates and calls for violence against Ukrainians based on their belonging to the

---

44 See UNGA Res 68/262, *Territorial integrity of Ukraine* (27 March 2014) UN Doc A/RES/68/262; and 'Treaty between the Russian 16 Federation and the Republic of Crimea on Admitting to the Russian Federation the Republic of Crimea and Establishing within the Russian Federation New Constituent Entities' (18 March 2014).

45 Vladimir Putin, 'Address by President of the Russian Federation' (The Kremlin, 18 March 2014) accessed on 15 May 2024, cited in FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (n 401) § 42; and *Ukraine v Russia* App nos 20958/14 and 38334/18 (ECtHR Grand Chamber 16 December 2020) § 324. This speech is one of the first noted references by the head of state to Ukrainians as Nazis, rhetoric which he has continued to employ since.

46 Svitlana Mazepa, *War, Hate, Propaganda and the Internet: A Dangerous Combination* (SpringerBriefs in Law 2024).

47 Such as FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41).

48 See Independent International Commission of Inquiry on Ukraine, 'Oral Update' (Human Rights Council, 54th Session, 25 September 2023) UN Doc A/HRC/54/CRP 3, which notes that 'some of the rhetoric transmitted in Russian state and other media may constitute incitement to genocide.'; see also Essex Human Rights Centre Blog: Digital Verification Unit assists United Nations Independent International Commission of Inquiry on Ukraine with reports on: Digital Verification Unit assists United Nations Independent (i) sexual and gender-based violence, and (ii) speech crimes Digital Verification (*Essex HRC Blog* 22 December 2023) <https://hrcessex.wordpress.com/2023/12/22/digital-verification-unit-assists-united-nations-independent-international-commission-of-inquiry-on-ukraine-with-reports-on-digital-verification-unit-assists-united-nations-independent/> accessed 26 June 2026.

Ukrainian nation' amounts to persecution as crime against humanity under Article 7(1)(h) of the Rome Statute.[49]

217. In his announcement of Russia's invasion of Ukraine, Putin specifically stated the purpose thereof to be 'to protect people who have been subjected to abuse and genocide by the Kyiv regime for eight years' further stating that this would be achieved through the 'demilitarisation and denazification of Ukraine'.[50] On 16 March 2022, the International Court of Justice (ICJ) responded to Ukraine's 'Request for the Indication of Provisional Measures' by ordering the suspension of Russian military operations.[51] The ICJ noted that there was no evidence substantiating the claims made by the Russian Federation that such a genocide was committed on Ukrainian territory against ethnic Russians or Russian speakers from 2014 to 2022.[52]

218. FIDH has issued a report noting that the use of this kind of 'politics of memory' rhetoric – in which the present is conflated with the past – sets the ideological foundation for the ongoing conflict in Ukraine. Using a technique of 'historical framing' the current war in Ukraine is 'linked as the continuation or repeat of the Soviet Union's liberation campaign from Nazi fascism during the Second World War, laying the ground for Ukraine's so-called "denazification"'.[53] Russian state-sponsored propagandists increasingly adopted this narrative after the annexation of Crimea in 2014, deliberately manipulating a historical traumatic narrative in order to justify current policies.[54] Statements made by Russian officials on the war in Ukraine based on the 'politics of memory' and 'historical framing' have frequently called for the need to 'liberate Ukraine from Nazis' and referred to 'Ukraine as an artificial entity that is historically comprised in large part of territories that "belong" to Russia. It is claimed that Ukrainians are really Russians, with a single history, mentality and culture'.[55]

219. FIDH has extensively documented, in the context of Russia's ongoing full-scale invasion of Ukraine, the frequent use of hate speech by Russian

49 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) Executive Summary, 2.

50 Vladimir Putin, 'Address of the President of the Russian Federation' (The Kremlin, 24 February 2022) <http://kremlin.ru/events/president/news/67843> accessed 25 June 2026.

51 *Allegations of Genocide under the Convention on the Prevention and Punishment of the Crime of Genocide* (*Ukraine v Russian Federation*) Provisional Measures Order (16 March 2022) ICJ Rep 211.

52 *Ukraine v Russian Federation* (Chapter 3, n 51) § 59.

53 Jade McGlynn, *Memory Makers: The Politics of the Past in Putin's Russia* (Bloomsbury Academic 2023) 51; FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) § 68.

54 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) § 69.

55 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) § 69.

politicians and public personalities (such as well-known media presenters), and this has cultivated an environment where ongoing mass violence against Ukrainians – including civilians – is normalised (and therefore encouraged as an illocution) due to the 'gravity, frequency, and reach of their statements'.[56]

220. The FIDH Article 15 communication highlights the responsibility of five prominent Russian public figures as direct perpetrators of persecution and asserts that a sixth individual is responsible either for ordering 'the commission of speech acts under Article 25(3)(b) of the Rome Statute or, at the very least, for failing to prevent them as a superior within the meaning of Article 28(2)'.[57] The communication maintains that these alleged perpetrators are 'only a representative group of a larger state-sponsored incubator of hatred'.[58] They were identified by FIDH because they have leveraged their professional and social status to repeatedly, continuously, and intentionally propagate rhetoric as a matter of State policy in order to incite hatred and violence against Ukrainians.

221. The communication explores in detail how hate speech perpetuated by Russian officials in the context of a widespread or systematic attack by Russian forces against the civilian population and infrastructure of Ukraine constitutes persecution as a crime against humanity as defined in Article 7(1)(h) of the Rome Statute.[59]

222. Since 2014, in Eastern Ukraine, the Russian military has engaged in significant efforts to cut off access of Ukrainian citizens to sources of information, using specially imported technological equipment to broadcast Russian propaganda and disseminate disinformation.[60] One example of this, is a widely cited Russian State TV news story which repeatedly showed a female refugee claiming that Ukrainian troops had executed a three-year old boy.[61] No evidence was ever provided for her account and the story, which became known as the

56 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) Executive Summary, 2 – see also the discussion on gravity in §§ 336–346.

57 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) Executive Summary, 2 – see also the discussion on the individual criminal responsibility of alleged perpetrators in §§ 174–322.

58 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) Executive Summary, 2.

59 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) see the legal assessment on hate speech as the crime against humanity of persecution §§ 113–128.

60 Mazepa (Chapter 3, n 46).

61 Kayleen Devlin and Olga Robinson, 'Ukraine Crisis: Is Russia Waging an Information War?' (*BBC News* 23 February 2022) <https://www.bbc.co.uk/news/60292915> accessed 25 June 2026.

'crucified boy', was later retracted.[62] Similar stories of Russian media portraying Ukraine in a negative light through the use of emotive and questionable content are not uncommon. The exact impact of propaganda and disinformation on the conflict is complex and difficult to quantify. It is clear, however, that the use of information warfare in this conflict can serve the purpose of exacerbating 'underlying social and political tensions by promoting divisive or extremist views' which risks creating a more polarised environment.[63]

223. The 2022 invasion itself was preceded by a widespread disinformation campaign, and more specifically a false flag operation that attempted to justify the full-scale invasion. Moreover, at the same time as FIDH has documented how disinformation campaigns and the systematic manipulation of information are routinely used as Russian operational tools in the assault on Ukraine,[64] there has been a consistent pattern of Russian attacks on hotels housing media professionals, humanitarians and other civilians in Ukraine.[65] An investigation conducted by Ukrainian non-governmental organisation Truth Hounds and Reporters Without Borders consider how media, commentary from politicians and other public figures, and disinformation campaigns on Russian Telegram channels justify strikes on hotels by framing them as military targets. The investigation highlights the prominent overlap among these statements and platforms which conflates foreign journalists with mercenaries and depicts the targeted hotels as their 'operational bases'.[66]

224. Disinformation campaigns justifying attacks on hotels deliberately portray the professionals inside as legitimate targets. Because of these attacks, media professionals have reconsidered their choice of hotels and adjusted their security plans. This has impacted their ability to safely and effectively conduct their work covering Russia's war in Ukraine. Beyond the direct impact of the attack itself, including the lives lost and mental and physical harm caused, this has broader consequences because it limits war reporting and the free spread of information beyond the conflict zone. It reduces the presence of independent actors who are able to document and observe any potential violations of international law. The use of disinformation to justify physical attacks on the media and humanitarians thereby further undermines capacity

62 Devlin and Robinson, BBC (Chapter 3, n 61).

63 Mazepa (Chapter 3, n 46).

64 Organisation for Economic Co-operation and Development (OECD), 'Disinformation and Russia's War of Aggression against Ukraine: Threats and Governance Responses' (03 November 2022).

65 Truth Hounds and Reporters Without Borders (RSF), 'Last Check-In: The Russian Strikes on Ukrainian Hotels Silencing the Press' (May 2025) 40.

66 Truth Hounds and RSF (Chapter 3, n 65) 43.

to combat future disinformation campaigns, as press freedom, a functioning independent media and the presence of observers are crucial tools in combatting the spread of deliberate falsehoods – especially in times of war.

2.3 Case Vignette: Israel and the Occupied Palestinian Territory (oPt)

225. On 7 October 2023 Hamas led an attack on Southern Israel, abducting 223 civilians, capturing 27 Israeli soldiers, and killing approximately 1,200 people – over 800 of them civilians.[67] In retaliation, Israel launched a full-scale military offensive on the Gaza Strip, resulting in the reported large-scale displacement of Gaza's population.[68] Although precise numbers are disputed, the Hamas-run Palestinian health authorities estimate that as of November 2025 more than 70,525 people were killed by Israel's ground and air campaign.[69] A significant proportion of those killed are reported to be civilians – including many women and children.[70] According to the Committee to Protect Journalists, the victims include at least 263 journalists and media professionals.[71]

226. Multiple proceedings are pending at the International Court of Justice relating to this conflict – an application of the Convention on the Prevention and Punishment of the Crime of Genocide in the Gaza Strip brought by South Africa against Israel as well as an application relating to the restrictions placed on United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA) operations.[72] A third concerning Alleged Breaches of Certain

---

67 Amnesty International, '"You Feel Like You Are Subhuman": Israel's Genocide Against Palestinians in Gaza' (2024) 12, 13.

68 Amnesty International (Chapter 3, n 67) 12.

69 The Economist, 'How Many People Have Died in Gaza' (*The Economist* 08 May 2025) <https://www.economist.com/interactive/middle-east-and-africa/2025/05/08/how-many-people-have-died-in-gaza> accessed 24 June 2026 – stating that although doubts about figures issued by Palestinian authorities persist, in previous conflict they have roughly matched numbers issued by Israel and the UN after conflicts ended.

70 Matthew Ghobrial Cockerill, 'Civilian Casualties in Gaza: Israel's Claims Don't Add Up' (*Action on Armed Violence* 28 October 2024) <https://aoav.org.uk/2024/casualties-in-gaza-israels-claims-of-50-combatant-deaths-dont-add-up-at-least-74-of-the-dead-are-civilians/> accessed 24 June 2026; and OHCHR, 'Six-Month Update Report on the Human Rights Situation in Gaza: 1 November 2023 to 30 April 2024' (*OHCHR* 8 November 2024).

71 Committee to Protect Journalists (CPJ), 'Journalist Casualties in the Israel-Gaza War' (*CPJ* 13 October 2023, updated 25 June 2026) <https://cpj.org/2023/10/journalist-casualties-in-the-israel-gaza-conflict/> accessed 27 June 2026.

72 *Application of the Convention on the Prevention and Punishment of the Crime of Genocide in the Gaza Strip* (*South Africa v Israel*) Application Instituting Proceedings (29 December 2023) ICJ; *Obligations of Israel in Relation to the Presence and Activities of the United Nations, other International Organizations and Third States in and in Relation to the Occupied Palestinian Territory* (Advisory Opinion) (22 October 2025) ICJ.

International Obligations in respect of the Occupied Palestinian Territory brought by Nicaragua against Germany did not find traction at the Court.[73] The human rights organisation Amnesty International has concluded that the acts committed by Israel in Gaza constitute prohibited acts under Article II (a), (b), and (c) of the Genocide Convention.[74]

227. For its part, the ICC has issued arrest warrants for three Hamas leaders and two Israeli government officials, including the sitting Prime Minister Benjamin Netanyahu, for alleged war crimes and crimes against humanity, but not for genocide.[75] While Israel is not a State Party to the Rome Statute, Palestine became a State Party in 2015. In issuing the arrest warrant against Israeli leaders – with Israel not being Party to the Rome Statute – the Court reiterated that when asserting jurisdiction on the basis of the territoriality principle, the nationality of the offender is irrelevant.

228. The Court's jurisdiction over the situation in Palestine, however, was contested by Israel pursuant to Article 19(2). Israel argued 'that Palestine does not possess the competences required under international law to be able to delegate territorial jurisdiction to the Court' and that 'its claim that Palestine is not a State on the territory of which the alleged conduct occurred is in itself sufficient to make it the sole State whose acceptance of jurisdiction is required.'[76] The Israeli challenge was dismissed as premature by PTC I.

2.3.1 The Role of Social Media during the Gaza Conflict – Alleged Incitement and Crimes

229. From the onset, this conflict has reverberated in social media, which has been used extensively by both Hamas and Israel. On 7 October 2023, Hamas used Telegram to stream its attack on Israel and, in some cases, militants took over their victims' accounts, livestreaming their captivity and torture.[77] Most of

---

73 *Alleged Breaches of Certain International Obligations in respect of the Occupied Palestinian Territory* (*Nicaragua v Germany*) Provisional Measures, Order of 30 April 2024, I.C.J. Reports 2024, 560, 569.

74 Amnesty International (Chapter 3, n 67) 238.

75 Karim AA Khan, 'Statement of ICC Prosecutor Karim AA Khan KC: Applications for Arrest Warrants in the Situation in the State of Palestine' (*OTP ICC* 20 May 2024) <https://www.icc-cpi.int/news/statement-icc-prosecutor-karim-aa-khan-kc-applications-arrest-warrants-situation-state> accessed 27 June 2026.

76 *Situation in the State of Palestine* (Decision on Israel's Challenge to the jurisdiction of the Court pursuant to article 19(2) of the Rome Statute) ICC-01/18-374-21-11-2004 1/8 PT (21 November 2024) §§ 11, 14.

77 Daniella Cheslow, 'In Gaza, a "Tipping Point" for Social Media at War' (*Politico* 16 October 2024) <https://www.politico.com/newsletters/digital-future-daily/2024/10/16/in-gaza-a-tipping-point-for-social-media-at-war-00184001> accessed 27 June 2026.

the (presumed) primary footage of Hamas atrocities against civilians (including sexual assault) was first shared privately on message apps like Telegram and WhatsApp and then posted more widely on social media platforms.[78]

230. Statements made by Israeli officials on X (formerly Twitter) and YouTube were cited in the South African application to the ICJ in the section on 'Expressions of Genocidal Intent [...]'.[79] A report by Amnesty International similarly cited amongst their sources 'an extensive range of visual and digital evidence, including satellite imagery, video footage and photographs posted on social media [...]'.[80]

231. In its report, Amnesty International argued that there is a long-standing pattern of dehumanising discourse, including social media content posted by Israeli officials, decision-makers and other influential public figures to legitimise and normalise violence against Palestinians. It contends that such discourse has not only existed and been promoted with impunity, but was already actively encouraged by Israeli authorities prior to 7 October 2023.[81] For example, in the context of a 2014 Israeli military offensive in Gaza nicknamed Operation Protective Edge, the Special Advisers of the Secretary-General on the Prevention of Genocide and the Responsibility to Protect expressed concern regarding the 'flagrant use of hate speech in the social media, particularly against the Palestinian population' and the dissemination of 'messages that could be dehumanising to the Palestinians' as well as highlighting alleged calls 'for the killing of members of this group'.[82]

232. Amnesty International alleges that, following 7 October 2023 and throughout the current conflict, 'there has been a significant escalation in the use of dehumanising, derogatory and racist language towards Palestinians by Israeli government officials, Knesset members and high-ranking military officials'.[83] Only a few weeks into the conflict, a collective of UN Special Mandate holders argued that there was 'discernibly genocidal and dehumanising rhetoric

78 Tamara Kharroub, 'Disinformation and Hate Speech on Social Media Contribute to Inciting War Crimes Against Gaza' (*Arab Center Washington DC* 13 October 2023) <https://arabcenterdc.org/resource/disinformation-and-hate-speech-on-social-media-contribute-to-inciting-war-crimes-against-gaza/> accessed 27 June 2026.

79 *South Africa v Israel*, Application Instituting Proceedings (Chapter 3, n 72) §§ 101–107.

80 Amnesty International (Chapter 3, n 67) 14.

81 Amnesty International (Chapter 3, n 67) 236.

82 UN, 'Statement by the Special Advisers of the Secretary-General on the Prevention of Genocide, Mr. Adama Dieng, and on the Responsibility to Protect, Ms. Jennifer Welsh, on the situation in Israel and in the Palestinian Occupied Territory of Gaza Strip' (24 July 2014).

83 Amnesty International (Chapter 3, n 67) 239.

coming from senior Israeli government officials'.[84] The UN Committee on the Elimination of Racial Discrimination also observed a 'sharp increase in racist hate speech and dehumanisation directed at Palestinians since October 7th'.[85] In addition, concern was expressed by the Commissioner-General of the UNRWA, who stated that the crisis in Gaza 'is compounded by dehumanising language'.[86] Israel and its allies, particularly the United States of America, have denied these allegations.

233. For its part, Hamas has allegedly repeatedly called for the destruction of Israel and its Charter envisages the obliteration of Israel.[87] In the wake of 7 October, its leadership declared that it would conduct 7 October 2023 style attacks repeatedly in pursuit of its goals.[88] Hamas officials have also reportedly acknowledged that there were 'faults' in the 7 October attack but argued that its fighters did not intentionally target civilians.[89]

234. The purpose of the current analysis is not to pronounce upon the factual veracity of these contested claims, but rather to explore the relevance of speech acts in this context, particularly insofar as they were issued on or amplified by social media or AI technology. While the use of social media as a tool of conflict in Israel and Palestine is not new, the role it has taken in the current conflict is unprecedented. This also holds true for its potential value as evidence in an international court. Indeed, the reliance on social media posts within and outside of judicial proceedings raises a multitude of questions regarding their proof of hate speech or incitement to genocide, their use in disinformation campaigns, and their impact on evidentiary processes like

84 UN OHCHR, 'Gaza: UN Experts call on international community to prevent genocide against the Palestinian people' (16 November 2023).

85 UN OHCRC, 'Occupied Gaza Strip: UN Committee calls for immediate ceasefire and urges end to hate speech amid fears of broader conflict' (27 October 2023).

86 UNRWA, 'The Gaza Strip: 100 days of death, destruction and displacement – Statement by Philippe Lazzarini, Commissioner-General of UNRWA' (13 January 2024).

87 Article 19, 'Israel and Palestine: A year on, the assault on freedom of expression continues', (7 October 2024) https://www.article19.org/resources/israel-and-palestine-a-year-on-the-assault-on-freedom-of-expression-continues/ accessed 30 June 2026; Memri TV, 7 (April 2023) https://www.memri.org/tv/rafah-gaza-friday-sermon-hamas-official-regeb-allah-necks-jews-annihilate-them accessed 30 June 2026.

88 NBC News, 'Hamas Official Claims it's Their "Legal Right" to Fight Against Israeli Occupation' (*NBC News* 03 November 2023) <https://www.nbcnews.com/now/video/hamas-official-vows-repeat-attacks-on-israel-this-is-our-legal-right-196999237771> accessed 27 June 2026.

89 Al-Jazeera, 'Hamas Says October 7 Attack Was a 'Necessary Step', Admits to 'Some Faults'' (*Al Jazeera* 21 January 2024) <https://www.aljazeera.com/news/2024/1/21/hamas-says-october-7-attack-was-a-necessary-step-admits-to-some-faults> accessed 24 June 2026.

gathering, submitting, and verifying. Moreover, AI and self-learning algorithms pose new challenges to trustworthiness of online sources and accountability.[90]

235. While some of these broader thematic issues are discussed more throughout the Guidelines as they are generally applicable, this sub-section will present a brief case study on the use of social media specifically in the armed conflict in Gaza. Hateful and violent content online has reportedly reached unprecedented levels and even a global dimension during the Gaza war.[91] The HRC found in a 2024 report that social media sites have provided fertile and ungoverned platforms for incitement to violence and hatred.[92] Underneath claim and counterclaim in the Gaza case there is a complex evidentiary and doctrinal legal question concerning when dehumanising, violent online rhetoric, combined with catastrophic military conduct, crosses the line into 'war crimes with hateful speech', and then further into genocide, with demonstrable specific intent to destroy a protected group in whole or in part.

2.3.2 Incitement and Potential Speech Crimes during the Gaza Conflict

236. Those asserting that Israeli officials have committed international crimes have referenced posts by Prime Minister Benajmin Netanyahu framing the conflict as a 'struggle between the children of light and children of darkness, between humanity and law of the jungle',[93] and an Old Testament reference characterising Palestinians as *amalek*. The term *amalek* has also been employed by Minister of Finance Bezalel Smotrich at a public event on 29 April 2024 (the week before an offensive on Rafah was launched) where he

90 Kharroub (Chapter 3, n 78); Amnesty International, 'Global: Social media companies must step up crisis response on Israel-Palestine as online hate and censorship proliferate' (27 October 2023) <https://www.amnesty.org/en/latest/news/2023/10/global-social-media-companies-must-step-up-crisis-response-on-israel-palestine-as-online-hate-and-censorship-proliferate/> accessed 26 June 2026.

91 Kharroub (Chapter 3, n 78); for the global impact see Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan, 'Global threats to freedom of expression arising from the conflict in Gaza' (23 August 2024) A/79/319.

92 HRC, 'Detailed findings on the military operations and attacks carried out in the Occupied Palestinian Territory from 7 October to 31 December 2023' (10 June 2024) A/HRC/56/CRP.4.

93 @IsraeliPM (Prime Minister of Israel), 'This is a struggle between the children of light and the children of darkness, between humanity and the law of the jungle' (Twitter/X, 16 October 2023, 4:07 pm GMT), <https://web.archive.org/web/20231017165958/https://twitter.com/IsraeliPM/status/1713949754948718657> accessed 27 June 2026.

reportedly called for the destruction of cities in Gaza, stating '[t]here are no jobs half done. Rafah, Deir al-Baleh, Nuseirat, destruction! Blot out the memory [of the people of] Amalek from under heaven'.[94] In its report, Amnesty International explains that the term *amalek* originates in the Hebrew Bible and refers to 'a nation who attacked the Israelites and were then, in an act of revenge, annihilated by them'.[95]

237. Accusations against Israel point to statements by Israeli officials discussing a second Nakba, the proposed transfer of the civilian population, and hints at the re-establishment of an Israeli settlement in the Gaza strip.[96] Section D of the ICJ Application Instituting Process (South Africa v. Israel) similarly points to speech acts such as, among others, Israeli President Herzog's statement that all women in Gaza are wives or sisters of Hamas murderers, which it argues suggests that there is no civilian population.[97] There is also reference to both media personas and regular civilians using similar language in posts which are allegedly glorifying Israel's attacks on civilians in Gaza, support the destruction of Gaza and which advocate for the use of violence against Palestinians.[98] There have also been reported incidents of Israeli Defense Forces (IDF) soldiers' social media accounts showing videos of themselves blowing up civilian property, posing with personal effects apparently found in houses, and similar acts that are alleged to glorify and encourage violence against civilians.[99]

2.3.3 Disinformation and the Use of AI

238. The line between inciting speech and disinformation (which is discussed previously) is often blurry because disinformation can be equally harmful when, for instance, it creates misconceptions or perpetuates deliberate falsehoods that intentionally fuel hate.[100] Disinformation on social media, similar

94 Amnesty International (Chapter 3, n 67) 15.

95 Amnesty International (Chapter 3, n 67) 9.

96 HRC, Detailed Findings oPt (Chapter 3, n 92) § 33, 152.

97 *South Africa v Israel*, Application Instituting Proceedings (Chapter 3, n 72) § 64.

98 Amnesty International (Chapter 3, n 67) 12; HRC, Detailed Findings OPT (Chapter 3, n 92) §§ 45–48.

99 The Times of Israel, 'IDF struggles to throttle endless feed of soldiers posting misdeeds on social media', (18 May 2026), https://www.timesofisrael.com/idf-struggles-to-throttle-endless-feed-of-soldiers-posting-misdeeds-on-social-media/ accessed 30 June 2026.

100 Ronan Ó Fathaigh, 'The Perils of Legally Defining Disinformation' (04 November 2021) 10 Internet Policy Review <https://policyreview.info/articles/analysis/perils-legally-defining-disinformation> accessed 27 June 2026; Tim Hwang, 'Dealing with Disinformation: Evaluating the Case for Amendment of Section 230 of the Communications Decency Act' in Nathaniel Persily and Joshua A Tucker (eds), *Social Media and Democracy* (Cambridge University Press 2020) 252, 254.

to hate speech, has reached unprecedented levels and is spreading at accelerated speed.[101] In comparison to the quantity of extremist speech on Facebook during the attacks that have been characterised as genocidal in Myanmar as addressed further in the case vignette above in Chapter 3, the situation in Gaza is an illustration of war propaganda during the new digital era due to the spread across different platforms, the involvement of both government and non-government accounts, and the use of new technologies (including AI) to create and promulgate content.[102] Notably, approximately twenty percent of the accounts participating in these conversations about the Gaza conflict are deemed to be fake.[103]

239. For instance, after the Israel Defence Forces took control of the Al-Shifa hospital in Gaza City, a seven minute long video was released on social media detailing the discovery of a collection of weapons and ammunition which was used to support the official Israeli narrative that Al-Shifa Hospital was being used as a command-and-control centre by Hamas.[104] However, the BBC noted that the initially released video had been edited and the nature of what was filmed was disputed.[105]

240. There have also been reports of Israeli disinformation campaigns targeting US lawmakers,[106] a statement by President Herzog about women in Gaza,[107] and a post by Prime Minister Netanyahu on X showing two burned Israeli children which was claimed to have been created by AI.[108] A University

---

101 Kharroub (Chapter 3, n 78).

102 Clarisa Nelu, 'Exploitation of Generative AI by Terrorist Groups' (*International Centre for Counter-Terrorism* 10 June 2024) <https://icct.nl/publication/exploitation-generative-ai-terrorist-groups> accessed 26 June 2024; David Gilbert, 'Here's How Violent Extremists Are Exploiting Generative AI Tools' (*Wired* 9 November 2023) <https://www.wired.com/story/generative-ai-terrorism-content/> accessed 27 June 2026.

103 Kharroub (Chapter 3, n 78).

104 New York Times 'A Tunnel Offers Clues to How Hamas Uses Gaza's Hospitals' (14 February 2024), https://www.nytimes.com/interactive/2024/02/12/world/middleeast/gaza-tunnel-israel-hamas.html accessed 30 June 2026. See also BBC Tracking latest Israeli military advances in Gaza City – BBC, 25 September 2025 https://www.bbc.co.uk/news/live/c8jmwmepe40t accessed 30 June 2026.

105 New York Times 'A Tunnel Offers Clues to How Hamas Uses Gaza's Hospitals' (14 February 2024).

106 Joseph Gedeon, 'Israel Targeted More Than 120 US Lawmakers in Disinformation Campaign' (*Politico* 05 June 2024) <https://www.politico.com/news/2024/06/05/israel-targeted-lawmakers-in-disinformation-campaign-00161906> accessed 24 June 2026.

107 *South Africa v Israel*, Application Instituting Proceedings (Chapter 3, n 72) 64.

108 Will Bedingfield, 'Generative AI Is Playing a Surprising Role in Israel-Hamas Disinformation' (*Wired* 30 October 2023) <https://www.wired.com/story/israel-hamas-war-generative-artificial-intelligence-disinformation/> accessed 27 June 2026.

of California Berkeley team examined the material and concluded that AI was not used but that through the allegations 'the seed of suspicion had already been planted'.[109] Hamas has reportedly used disinformation tactics to generate support for its campaign.[110] An AI-generated picture showed a Palestinian father with his five children in front of a pile of rubble.[111] The picture was shared many times on social media and allegedly showed the inhumanity of Israel's bombing of civilians. Only a few days later, an analysis revealed that it was fake, by pointing to numerous errors and inconsistencies that are typical for AI images (e.g. there are too many or too few fingers and toes in several of the hands and feet in the picture).[112]

241. These examples illustrate how AI has become a potential tool for creating material used in disinformation campaigns. AI, further, exerts influence over algorithms that encourage engagement which often leads to the amplification of harmful and inflammatory content in a disproportionate manner.[113] It also makes fact-checking and evidential verification more difficult.[114] During periods of intense conflict, parallel information wars by both sides – and in particular the dissemination of disinformation – carry a strong risk of changing the tide of public opinion and contributing to an environment conducive to the commission of international crimes.[115]

242. The conflict in Gaza serves as a valuable case study of inciting speech on social media. The conflict is not only fought on the ground but also on online platforms, fuelling, through algorithms and AI, rapid dissemination of hateful content. Furthermore, AI plays a key role in disinformation campaigns that create misconceptions about the enemy and justify cruel methods of war. It remains to be seen whether social media can play a role in mitigating inciting

109 Bedingfeld (Chapter 3, n 108).

110 Armin Langer, 'Antisemitic Myths on the Web Amidst Global Crises of the Early 2020s' in Florian Hartleb (ed), *Antisemitism on the Rise: New Ideological Dynamics* (European Institute for Counter Terrorism and Conflict Prevention 2024) 71, 73; Elizabeth Dwoskin, 'Growing Oct 7 'Truthers' Say Hamas Massacre Was a False Flag' (*Washington Post* 21 January 2024) <https://www.washingtonpost.com/technology/2024/01/21/hamas-attack-october-7-conspiracy-israel/> accessed 27 June 2026.

111 Ines Eisele and Uta Steinwehr, 'Fact check: AI fakes in Israel's war against Hamas' (*Deutsche Welle* 11 October 2023) <https://www.dw.com/en/fact-check-ai-fakes-in-israels-war-against-hamas/a-67367744> accessed 26 June 2026.

112 Eisele and Steinwehr (Chapter 3, n 111).

113 Amnesty International (Chapter 3, n 67) 12.

114 Kharroub (Chapter 3, n 78) 8.

115 New York Times 'A Tunnel Offers Clues to How Hamas Uses Gaza's Hospitals' (14 February 2024); Joseph Gedeon, 'Israel Targeted More Than 120 US Lawmakers in Disinformation Campaign' (5 June 2024).

speech and disinformation campaigns and how these posts and videos can serve as possible evidence in international criminal proceedings. Given the increasingly ubiquitous role of social media in armed conflicts and the potential commission of atrocities, it is apposite to look to the potential accountability for these outcomes under international law.

3 Pathways to Liability for Social Media-Driven International Crimes

3.1 Corporate Criminal Liability for International Crimes

243. Turning to the question of who may be held responsible for the perpetration of speech crimes on social media, a key issue is corporate criminal liability. The case studies in the previous sub-sections highlight the growing concerns about the different ways in which social media platforms become sites for actors to encourage the commission of international crimes. Having explored the potential role of social media in the commission of international crimes, the following section explores the possible pathways to liability that might exist for social media corporations and executives.

244. The starting point for any discussion of accountability for harm in the digital sphere under international criminal law is Article 25(1) of the Rome Statute which establishes that the ICC's jurisdiction includes only 'natural' persons and as such does not extend to 'legal persons' in the broader sense.[116] As such, there is no legal basis for corporate criminal liability under the ICC framework. Whether jurisdiction over legal persons *should* be included – and as such whether corporations would have criminal liability – was considered

116 Rome Statute (Key Guidelines, n 2) art 25(1). As discussed in Juan Pablo Calderon Meza, 'ICC Personal Jurisdiction on Corporations for Criminal Liability and/or Civil Liability for Reparations' (*Harvard International Law Journal Blog* 13 May 2021) <https://journals.law.harvard.edu/ilj/2021/05/icc-personal-jurisdiction-on-corporations-for-criminal-liability-and-or-civil-liability-for-reparations/> accessed 27 June 2026, the travaux préparatoires included personal jurisdiction over legal persons, other than States 'when the crimes committed were committed on behalf of such legal persons or by their agencies or representatives'. See Preparatory Committee on the Establishment of an International Criminal Court, 'Text of the Draft Statute for the International Criminal Court – Part 3. General Principles of Criminal Law' (01 April 1998) A/AC.249/1998/CRP.9 art 17(5)–(6); see also Committee of the Whole Working Group on General Principles of Criminal Law, 'Working Paper on Article 23, Paragraphs 5 and 6' (03 July 1998) A/CONF.183/C.1/WGGP/L.5/Rev.2 art 23(5)–(6).

during the drafting of the Rome Statute.[117] Complementarity emerged as the main point of contention since some States Parties' legal systems do not have mechanisms to pursue corporate criminal liability.[118]

245. The focus on individual criminal accountability for acts undertaken in association with a corporation has been the approach followed by the ad-hoc tribunals. A notable exception is the approach of the Appeals Panel of the Special Tribunal for Lebanon in the contempt cases of *New TV S.A.L. and Karma Mohamed Tahsin Al Khayat* and *Akhbar Beirut S.A.L. Ibrahim Mohamed Ali Al Amin* which closely followed *New TV S.A.L. and Al Khayat*.[119] These decisions in two separate contempt cases dealt with the names of confidential tribunal witnesses published by media companies (a television station and a newspaper, respectively). New TV S.A.L. was further accused of having violated a court order for separately failing to remove the names from their website and YouTube channel.

246. In both cases, charges were imputed against a natural person (the individual concerned) alongside the legal person (the corporation). On the charges against the corporation, the defence teams challenged the STL's jurisdiction over legal persons. In the *New TV S.A.L. and Al Khayat* decision, the Contempt Judge Lettieri first found that the Tribunal lacked such jurisdiction over legal persons – holding that rule 60*bis* of the STL Rules of Procedure and Evidence (STL Rules) applied only to natural persons.

247. However, the majority on the Appeals Panel subsequently disagreed with this decision, finding that there exists sufficient ambiguity in the meaning of 'person' under Rule 60*bis* of the STL Rules to contemplate a prosecution

117 UN Diplomatic Conference of Plenipotentiaries on the Establishment of an International Criminal Court, 'Official Records Volume II (15 June–17 July 1998) AA/CONF.183 (Vol. II); See also Alessandra De Tommaso, *Corporate Liability and International Criminal Law* (Routledge, Taylor & Francis Group 2024) 83–87.

118 William Schabas, *The International Criminal Court – A Commentary on the Rome Statute* (Oxford University Press 2016) 564–566; Jaya Bordeleau-Cass, 'The "Accountability Gap": Holding Corporations Liable for International Crimes' (2019) 3 PKI Global Justice Journal <https://globaljustice.queenslaw.ca/news/the-accountability-gap-holding-corporations-liable-for-international-crimes#:~:text=As%20it%20is%20currently%20framed,liable%20through%20international%20criminal%20law> accessed 26 June 2026; Elies van Sliedregt, 'The Future of International Criminal Justice Is Corporate' (2015) 23:1 Journal of International Criminal Justice 17.

119 *Prosecutor v. Al Jadeed S.A.L. and Karma Mohamed Tahsin Al Khayat* (Interlocutory Appeal Against the Decision on Motion Challenging Jurisdiction) STL-14-05/PT/AP/AR126.1 (31 July 2014); and *Prosecutor v Akhbar Beirut S.A.L. and Ibrahim Mohamed Ali Al Amin* (Decision on Interlocutory Appeal Concerning Personal Jurisdiction in Contempt Proceedings) STL-14-06/PT/AP/AR126.1 (23 January 2015).

of the New TV S.A.L. corporation alongside Ms Al Khayat. They justified this approach through statutory interpretation by reference to Rule 3 of the STL Rules,[120] which they held required an interpretation of the rules 'in a manner consistent with the spirit of the Statute rather than the (black) letter of the law'.[121]

248. Notably, given that the STL is a 'hybrid court' integrating domestic and international criminal law, this conclusion was also supported by Lebanese domestic law which allows for legal persons (in this case, a television media corporation) to be held criminally liable.[122] In its reasoning, the Panel further makes reference to other 'civil law countries which have adopted corporate criminal responsibility'[123] and states that 'criminal responsibility of, and jurisdiction over legal persons has been widely adopted as essential to closing an impunity gap'.[124]

249. In the *New TV S.A.L. and Al Khayat* case, the Appeals Panel thus found that the Contempt Judge had committed an error in disregarding relevant national law and major international trends recognising the criminal liability of legal persons. The Appeals Panel thus stated that:

> The global reality is that legal persons (and in particular, corporations) have become powerful actors in every sector of modem life, often dwarfing natural persons and playing dominant roles in our societies, for both good and bad. In this context, the vast majority of national legal systems have recognised the need to close the gap of corporate impunity and hold legal persons, as well as natural persons, criminally responsible. The overwhelming trend is to extend criminal liability to legal persons.[125]

250. The flexibility applied by the Appeals Panel in interpreting the STL Rules and Procedures and thereby including legal persons within the jurisdiction for

120 Manuel J Ventura, 'The Prosecution of Corporations before a Hybrid International Criminal Tribunal' (2016) 2:1 African Journal of International Criminal Justice 71, 74 footnote 6; Special Tribunal for Lebanon (STL), 'Rules of Procedure and Evidence' (adopted on 20 March 2009, last amended on 10 April 2019) STL-BD-2009-01-Rev.10 Rule 3 Interpretation of the Rules, 19.

121 Ventura (Chapter 3, n 120) 74.

122 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 30 states that '[t]he fact that Lebanese law clearly includes corporate criminal liability and that jurisdiction over legal persons was not expressly excluded from the Tribunal Statute demonstrates the intention of the drafters not to exclude it. There is certainly no reason to exclude it from the Tribunal's inherent powers.'

123 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 30.

124 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 33.

125 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 27.

contempt should be understood within the context of the Lebanese domestic legal system, which has adopted corporate criminal responsibility.[126] The Panel also cites the fact 'that jurisdiction over legal persons was not expressly excluded from the Tribunal Statute demonstrates the intention of the drafters not to exclude it'.[127] In highlighting the need to fill the gap in accountability for corporations that violate orders by international courts, the Appeals Panel cites from the *Ayyash et al.* Appeals case:

> the tenet of construction that a statute is presumed to be 'always speaking' recognizes the reality that society alters over time and interpretation of the law may always evolve to keep pace. The power and capabilities of corporations, in terms of the assets that they can bring to bear, the harm they can do and the ways they can shield individual natural persons in the corporate machinery from accountability are realities that international courts cannot ignore, but must address.[128]

251. The Judges in the *New TV S.A.L. and Al Khayat* Appeals Panel also opined that too narrow an interpretation of the Court's 'important power to administer justice and police its proceedings' – while not rendering that power entirely 'meaningless', creates an array of challenges to its effective operation.[129]

252. Following the finding by the Appeals Panel in *New TV S.A.L. and Al Khayat*, the separate case of *Akhbar Beirut S.A.L. Al Amin* with similar facts – and presenting the same jurisdictional challenge – came before the same Contempt Judge (Lettieri) who had initially dismissed the charges against the legal person in *New TV S.A.L. and Al Khayat.*[130] Remarkably, the Judge Lettieri repudiated the findings of the Appeals Panel in *New TV S.A.L. and Al Khayat* 'noting, among other things, that the Appeals Panel had exclusive jurisdiction over that case alone and therefore its legal holdings did not extend to the separate *Akhbar Beirut S.A.L.* case'.[131] When this case then went again before an Appeals Panel, the Judge's determination was the same as in the *New TV S.A.L.*

---

126 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 30.

127 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 30.

128 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 34 citing *Prosecutor v Salim Jamil Ayyash, Mustafa Amine Badreddine, Hussein Hassan Oneissi, and Assad Hassan Sabra* (Interlocutory Decision on the Applicable Law: Terrorism, Conspiracy, Homicide, Perpetration, Cumulative Charging) STL-11-01i, Interlocutory Decision on the Applicable Law: Terrorism, Conspiracy, Homicide, Perpetration, Cumulative Charging (I 6 February 2011) § 20 Applicable Law.

129 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 37.

130 Ventura (Chapter 3, n 120) 77.

131 Ventura (Chapter 3, n 120) 77.

and Al Khayat case – they again held that the STL's jurisdiction also included legal persons.

253. Although the *New TV S.A.L. and Al Khayat* and *Akhbar Beirut S.A.L. Al Amin* cases are not concerned with the core international crimes, the reasoning of the Judges in these cases is nevertheless germane because they represent the first examples of corporations being prosecuted before an international tribunal – albeit one which is hybrid in character and for the crime of contempt. As such, the decisions in these cases 'signify a landmark moment and perhaps a pivotal legal development that could have wide repercussions'[132] and one which is of particular relevance to the emerging discussion on potential avenues towards accountability for the potential role played by social media in the perpetration of speech crimes.

254. Indeed, the possible challenges towards such accountability are predicted in *New TV S.A.L. and Al Khayat*, where referring back to the circumstances of the case, the Appeals Panel, looking at the role and realities specifically of media legal persons, states that the large-scale publication of information and widespread dissemination thereof to the population of an entire country is, 'in most instances, not accomplished by the work of one or two natural persons, but by the very substantial apparatus, resources and workings of a corporation or similar entity operating newspapers, television stations, internet sites and other media outlets'.[133]

255. Consequently, 'given the operation of complex corporate machinery' identifying the most responsible individuals becomes a very difficult (or even impossible) task and the '*actus reus* and *mens rea* elements may be spread among multiple persons at various corporate levels such that no natural person can be effectively prosecuted'.[134] In such situations, senior corporate managers may hide behind the corporate entity itself and delegate to lower-level employees the carrying-out of particular steps in an alleged crime.[135] It may therefore not always be possible nor effective to prosecute 'each cog in the wheel'.[136] Furthermore, the Appeals Panel highlights the risk of creating 'a zone of impunity for legal persons and indirectly for natural persons, too, where, as indicated above, they can hide among the corporate machinery'.[137]

---

132 Ventura (Chapter 3, n 120) 79.
133 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 35.
134 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 36.
135 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 36.
136 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 36.
137 *New TV S.A.L. and Al Khayat* AP (Chapter 3, n 119) § 37.

256. Currently the ICC does not have jurisdiction to adjudicate corporate criminal liability, although there are arguments which can be made both in favour and against it. The increasing involvement of national jurisdictions in prosecuting atrocity crimes via universal jurisdiction raises the possibility of imposing such liability, including for social media amplified or facilitated speech crimes. The insights from the STL's cases addressing legal persons provide guidance by analogy regarding the application of key legal tests to corporate structures, particularly the controlling mind test and other legal mechanisms to establish the culpability of non-natural legal persons.

257. In the digital and AI age, there are several arguments in favour of corporate criminal liability for speech crimes.[138] Collective action by a powerful corporate entity can result in a greater harm than individual action. When a wrong has been committed in the aggregate, for example by channelling it through a corporation, it may be arduous to identify a single person liable under international criminal law because the Prosecution cannot establish a defined causal link for individual actions. Furthermore, pursuing corporate liability does not preclude individual liability but rather could add a potentially effective means of deterrence for collective actions, which require systemic punishment rather than merely imposing ad hoc liability for certain individuals who acted on behalf of the collective.[139] The historical argument that corporate liability should be excluded from international criminal law because of a lack of consensus within national jurisdictions requires revisiting in the current era.[140]

258. However, there are also several arguments which can be raised against liability for corporations and social media companies in particular. Criminalising corporations (or even individuals) could lead to arbitrary and overboard censoring of users including 'shadow-banning' on social media platforms.[141] Shadow banning refers to a form of covert moderation on social media or online platforms where a user's posts or activity become less visible or entirely hidden – often without any notification from the company. The experience of being shadow banned can be distressing and also risks removing a vital tool for communication and documentation in crises and conflicts both because of

138 Ronald C Slye, 'Corporations, Veils, and International Criminal Liability' (2008) 33 Brooklyn Journal of International Law 955, 959.

139 Slye (Chapter 3, n 138) 962.

140 David Scheffler, 'Corporate Liability under the Rome Statute' (2016) 57 Harvard International Law Journal 35, 38.

141 Hamilton (Chapter 3, n 2) 221.

the actual shadow-banning practice and possible self-censoring behaviour.[142] There is little transparency around the practice of 'shadow-banning', which has not been formally acknowledged by social media corporations like Meta.[143]

259. Although shadow-banning does not fit neatly into the category of speech crimes and is rather concerned with censorship and the suppression of speech, if such practices do take place, they have the potential to shape discourse and influence public opinion. Because the content moderation rules of social-media companies are already far reaching (often going beyond those of States) and apply a wide definition of hate speech, it is necessary to take into account the way in which these policies and rules are applied and the power this ascribes to social media corporations to regulate information and content. While harmful content should be subject to moderation, the label of speech crimes should not be used in an overly broad manner as a justification for suppressing speech or excusing censorship. To avoid encroaching on individual liberties and chilling free speech, speech crimes under international law must be precisely defined and subjected to requirements such as the gravity threshold at the ICC.

3.2 Social Media Algorithms and Liability

260. If it is impossible or undesirable to hold corporations criminally liable for content regulation, can there be liability for the algorithm used to suggest content, and elevate or depress users' posts?[144] Corporate culpability for algorithms has not been tested in international criminal courts. A number of US cases, however, have examined this question under Section 230(c)(1) of the US Communications Decency Act (CDA, 1996), which provides online social media platforms with publisher immunity. Section 230 of the CDA allows the hosts of online platforms to moderate and remove harmful content without the accompanying legal responsibilities of publishers or speakers (e.g. to refrain from making true threats, revealing confidential state secrets etc.). It states that '[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another

142 Deborah Brown and Rasha Younes, 'Meta's Broken Promises: Systemic Censorship of Palestine Content on Instagram and Facebook' (*Human Rights Watch* 21 December 2023) <https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-%20instagram-and> accessed 26 June 2026.

143 Brown and Younes, HRW (Chapter 3, n 142).

144 Eric N Holmes, 'Liability for Algorithmic Recommendations' (*Congressional Research Service* 2023) <https://www.congress.gov/crs-product/R47753> accessed 26 June 2026.

information content provider'.[145] Although there are some exceptions for federal crimes, violations of intellectual property, and sex trafficking, the approach to concerns related to the spread of disinformation and hate speech in online spaces remains contested through Section 230 CDA.[146] Precedent-setting US cases which have considered the question of liability of online platforms for harmful content promoted by their algorithms include *Force v Facebook, Inc.*; *Twitter, Inc. v Taamneh*; and *Gonzalez v Google LLC*.[147]

261. In *Force v Facebook, Inc.*, 934 F.3d 53 (2nd Cir. 2019) the relatives of victims of a terrorist attack alleged that Facebook had unlawfully assisted a terrorist organisation (Hamas) by allowing it to publish content and recruit on its social media platform and 'mak[ing] that content more visible, available, and usable'.[148] They contended that Facebook's use of algorithms to display personalised friend and content suggestions extended its role beyond that covered within the scope of Section 230 because in doing so, it actively contributed to the development of said content.[149] The US Court of Appeals for the Second Circuit Court rejected the claim that algorithmic sorting or recommendations of certain content contributed materially to the development of content.[150] This ruling essentially recognised algorithmic recommender systems as automated and neutral tools. The Supreme Court declined to hear an appeal of the case in 2020, although Chief Judge Robert Katzman stated in his dissenting opinion that, 'mounting evidence suggests that providers designed their algorithms to drive users toward content and people the users agreed with – and that they have done it too well, nudging susceptible souls ever further down dark paths'.[151] The case of *Force v Facebook, Inc.* is significant because it affirmed that Section 230 immunity applied even in a terrorism related case, which solidified broad protections for social media companies by affirming that their platforms should be treated as publishers.

262. In *Twitter v. Taamneh*, the US Supreme Court considered whether the social media company could be held responsible for aiding and abetting a

145 Valerie C Brannon and Eric N Holmes, 'Communications Act of 1934 Section 230' (*Congressional Research Service* 2024) <https://www.congress.gov/crs-product/R46751> accessed 26 June 2026.

146 Brannon and Holmes (Chapter 3, n 145).

147 For an in-depth discussion of liability for algorithmic recommendations in the US context, see Holmes (Chapter 3, n 144).

148 *Force v Facebook, Inc* 934 F.3d (31 July 2019) 49.

149 *Force v Facebook* (Chapter 3, n 148) 45–48.

150 *Force v Facebook* (Chapter 3, n 148) 45–48.

151 *Force v Facebook* (Chapter 3, n 148) Katzmann, Chief Judge, concurring in part and dissenting in part 33.

terrorist attack at the Reina nightclub in Istanbul in 2017 that killed 39 people when its 'recommendation' algorithms promoted content posted by Islamic State of Iraq and Syria (ISIS).[152] ISIS used Twitter,[153] YouTube, and Facebook to recruit, fundraise, and spread propaganda and the platforms' recommendation algorithms amplified ISIS content. The plaintiffs alleged that the social media companies knew ISIS used their platforms but failed to remove all ISIS-related accounts, and thus aided and abetted (under 18 U.S.C. § 2333(d)-Civil Remedies) ISIS in the Reina nightclub attack.

263. Justice Clarence Thomas writing in 2023 for the (unanimous) majority expressed scepticism about the *actus reus* claims about Twitter's algorithm advanced by the plaintiffs and reversed the Judgement of the Ninth Circuit. As opposed to distinctly promoting ISIS content (or locutions), the Court found that the algorithms were 'agnostic as to the nature of the content' and therefore Twitter could not be held liable for utilising 'passive aids' to engage in content distribution.[154]

264. Moreover, the intent requirements were not met. The plaintiffs' theory of aiding-and-abetting liability failed because they did not plausibly allege that Twitter (or the other platforms) 'knowingly' and substantially assisted the Reina nightclub attack and did not represent the kind of affirmative assistance required under the Justice Against Sponsors of Terrorism Act (JASTA): 'The point of aiding and abetting is to impose liability on those who consciously and culpably participated in the tort at issue. The focus must remain on assistance to the tort for which plaintiffs seek to impose liability.'[155]

265. Yet the Supreme Court made its finding on a narrow basis and did not seek to generalise its decision. It emphasised that Twitter's immunity stemmed from the absence of a nexus between the algorithm's promotions and the terrorist attack that had affected the plaintiffs.[156] This leaves open the possibility that a platform could incur liability if such a nexus were proven by future plaintiffs and a platform acted knowingly to advance content that incited mass crimes. In her concurrence, Justice Ketanji Brown Jackson left the door open to future cases in noting that the Court's finding in this case 'do not necessarily translate to other contexts', and '… the Court's view of the facts – including its characterisations of the social-media platforms and algorithms at

152 *Twitter, Inc v Taamneh et al.*, 598 US 471 (2023).

153 The platform's name was rebranded from Twitter to X in July 2023. The acts considered in *Twitter v. Taamneh* occurred prior to this name change.

154 *Twitter v Taamneh* (Chapter 3, n 152) 23.

155 *Twitter v. Taamneh* (Chapter 3, n 152) 30.

156 *Twitter v Taamneh* (Chapter 3, n 152) 30.

issue – properly rests on the particular allegations in those complaints. Other cases presenting different allegations and different records may lead to different conclusions.'

266. In *Gonzalez v. Google LLC*, 598 US 617 (2023) a US exchange student was killed along with 129 other people during an ISIS coordinated terrorist attack in Paris in 2015. The plaintiffs sought legal remedy against Google as parent company of YouTube, arguing that the company is partially responsible for her death because the algorithm tailored content based on user profiles which led users towards recruitment videos for ISIS. The plaintiffs further alleged that Google's recommendation algorithms amplified ISIS content and that Google shared advertising revenue with ISIS (a claim Google disputed). There was no allegation that Google knew about, or assisted in, the Paris attacks.

267. Google relied on Section 230 for its defence, which provides immunity from content published on an internet service provider's platform by third-party users. A lower court ruled in favour of Google, the lower court's decision was upheld by the Ninth Circuit Court of Appeals and the family then appealed to the Supreme Court which combined the case with *Twitter v. Taamneh*. In the appeal to the Supreme Court, the family focused on the fact that the YouTube algorithm had been tailored to deliver content that would attract the attention of the end user – which even if automatically done, they argue is a form of content moderation not fully covered by Section 230. When the Court ruled unanimously in *Twitter* that aiding and abetting requires proof of knowing participation and substantial assistance in a specific act of terrorism, *Gonzalez* was sent back to the lower courts on a per curiam decision with instructions to consider the Court's decision in *Twitter*.[157]

3.3 Liability of Natural Persons for Social Media Corporations

268. The case studies presented in the previous sections highlight how courts have struggled to regulate instances where individuals or groups use social media platforms to encourage and contribute to the commission of international crimes. The current lack of corporate liability within the international criminal law framework leaves open the accountability gap for speech crimes on social media and other online platforms. At the domestic level, the question of liability for algorithmic recommender systems has become particularly contentious. This sub-section builds on this further by exploring what pathways and challenges towards responsibility and accountability might exist regarding the liability of executives of social media executives.

157 *Reynaldo Gonzalez et al., Petitioners v Google LLC* 598 US 617 (2023).

269. Although it is not specifically concerned with social media and technology entities, some comparisons can be drawn from the existing international criminal law jurisprudence which offer insights into the overall issue of liability of individual natural persons for their activities within corporate structures. This may represent a promising approach to the question of criminal liability of social media executives.

270. These comparisons rely primarily on two distinct strands of argumentation. One draws a parallel between social media companies and news outlets, such as radio or TV stations that provide a platform through which to broadcast and disseminate messaging and relies on the IMT *Fritzsche* and the ICTR *Nahimana* cases as precedents. A second frames social media companies as analogous to weapons suppliers and references the post-World War II *Zyklon B* trial and the *Taylor* case at the SCSL. Additionally, in the Cote d'Ivoire case of Laurent Gbagbo and Charles Blé Goudé at the ICC, the latter's speeches were alleged by the Prosecution to have contributed to crimes. However, in its no-case-to-answer decision, the Trial Chamber concluded that the Prosecution had failed to demonstrate the causal nexus between his speeches and any charged crimes.[158] This illustrates again the need to clearly demonstrate the causal nexus when speech is charged as a mode of liability.

271. This section presents an overview of relevant cases before drawing out parallels that isolate potential pathways towards liability for social media executives. It also considers the complications that may emerge. Considering the increasing use of social media as a site for the commission of speech crimes across the globe, as evidenced in the various case studies presented thus far, this is a subject growing in importance. It raises significant questions regarding the lack of accountability for the corporations of the digital era which are currently among the largest corporations in the world by market capitalisation, dwarfing the gross domestic product of many nation-states.

3.3.1 Case Vignette: Nahimana

272. In the *Media Case*, which has been detailed above, Ferdinand Nahimana was charged with direct and public incitement to genocide on the basis of his

158 'Requête de la Défense de Laurent Gbagbo afin qu'un jugement d'acquittement portant sur toutes les charges soit prononcé en faveur de Laurent Gbagbo et que sa mise en liberté immédiate soit ordonnée': *The Prosecutor v. Laurent Gbagbo and Charles Blé Goudé* (Trial Chamber Reasons of Judge Geoffrey Henderson) ICC-02/11-01/15-1263 (2019) §§ 2007–2027.

speeches and broadcasts at the radio station RTLM before and after April 6, 1994 – the day mass killings began.[159] During these broadcasts, he called for the killing of specific individuals. According to the ICTR Trial Chamber, RTLM, also dubbed 'Radio Machete', was its editor's 'weapon of choice … used to instigate the killing of Tutsi civilians.'[160] Nahimana and his two co-defendants, Jean-Bosco Barayagwiza (a lawyer and RTLM executive) and Hassan Ngeze (founder, owner, and editor of the newspaper *Kangura*), were convicted at the ICTR, *inter alia*, of direct and public incitement to genocide, conspiracy, instigating genocide, extermination, and persecution.[161] The Appeals Chamber largely upheld the Trial Chamber's convictions, setting aside those charges that related to speech prior to the start of the genocide.[162]

273. The Trial Chamber in *Nahimana* found that 'editors and publishers have generally been held responsible for the media they control.'[163] The went on, in a surprisingly direct fashion, to say that publishers and editors can be regarded as equally responsible as the inciters themselves because they provide a forum for criminal speech, and that owners of media outlets have 'the power to shape the editorial direction.'[164] They stated that 'whether or not the purpose in publicly transmitting the material was of a bona fide nature' would be the test to establish criminal liability.[165] The Trial Chamber relied on various arguments to establish Nahimana's control over RTLM: (1) he and Barayagwiza effectively controlled the management of RTLM, (2) he was seen as the founder and director of RTLM, (3) he and Barayagwiza represented the radio station externally in official capacity, (4) he controlled financial operations, (5) he held supervisory responsibility for all activities and could take remedial action when necessary, (6) Nahimana played an active role in determining the content of RTLM broadcasts, (7) after April 6, 1994 he continued to have *de jure* authority, and (8) his intervention in early July of 1994 to stop the broadcasting attacks on General Dallaire and the United Nations Observer Mission Uganda-Rwanda (UNAMIR) was seen to be an indicator of his continuing *de facto* control over RTLM.[166]

---

159 Wilson and Gillett (Key Guidelines, n 1) § 88; Catharine A MacKinnon, 'Prosecutor v. Nahimana, Barayagwiza, & Ngeze Case No. ICTR 99–52-A' (2009) 103 American Journal of International Law 97, 97.

160 *Nahimana et al.* TC (Introduction, n 2) §§ 342, 1031.

161 *Nahimana et al.* TC (Introduction, n 2) §§ 947, 977, 1055.

162 MacKinnon (Chapter 3, n 159) 97.

163 *Nahimana et al.* TC (Introduction, n 2) § 1001.

164 *Nahimana et al.* TC (Introduction, n 2) § 1001.

165 *Nahimana et al.* TC (Introduction, n 2) § 1001.

166 *Nahimana et al.* TC (Introduction, n 2) §§ 970–1081.

274. While the Appeals Chamber overturned Nahimana's conviction for instigating genocide (Article 6 (1) ICTR-Statute) because of a lack of contemporaneousness, it upheld the conviction on direct and public incitement to genocide (Article 6 (3) ICTR-Statute) after April 6, 1994.[167] In overturning the conviction relating to Article 6(1) ICTR-Statute the Appeals Chamber stated that 'the role of founder of RTLM does not in itself sufficiently establish that the Appellant substantially contributed to the commission of the crime of genocide.'[168] However, the Appeals Chamber concluded that the Trial Chamber used the superior responsibility standard to establish Nahimana's criminal liability under Article 6(3) ICTR-Statute.[169]

275. Under the superior responsibility mode of liability, three elements had to be fulfilled. Nahimana had to have (1) effective control over his subordinates, (2) he had to know or had to have reasons to know that crimes were being committed, and (3) he had to have failed to take the necessary and reasonable measures to prevent such acts or to punish the perpetrators thereof.[170] As to the effective control requirement, the Appeals Chamber held that Nahimana had 'failed to show that RTLM came under control of the military after 6 April 1994.'[171] Consequently, it found that he 'continued to possess the power to intervene at RTLM.'[172] Like the Trial Chamber, the Appeals Chamber held that his (late) intervention with RTLM journalists to stop the attacks on General Dallaire and UNAMIR constituted *de facto* control.[173]

276. Regarding the *mens rea* requirement of specific intent to commit genocide, the Appeals Chamber held that Nahimana, as a member of the provisional board, knew that RTLM programming was generating concern. However, he defended RTLM programming at meetings with the Ministry of Information in 1993 and 1994, and RTLM programming followed its trajectory, steadily increasing in vehemence after 6 April 1994. The Appeals Chamber concluded that he 'knew or had reasons to know' that his subordinates had committed or were about to commit crimes.[174] As to the defendant's failure to take the necessary and reasonable measures, the Appeals Chamber concluded that even if he shared power with others that fact could not exonerate him and, further, that Nahimana had sufficient authority to take necessary and reasonable measures

167 *Nahimana et al.* (Appeal Chamber Judgement) ICTR-99-52-A (28 November 2007) § 857.
168 *Nahimana et al.* AC (Chapter 3, n 167) § 549.
169 *Nahimana et al.* AC (Chapter 3, n 167) § 777–857.
170 ICTR Statute (Chapter 2, n 195) art 6(3).
171 *Nahimana et al.* AC (Chapter 3, n 167) § 827.
172 *Nahimana et al.* AC (Chapter 3, n 167) § 827.
173 *Nahimana et al.* AC (Chapter 3, n 167) § 834.
174 *Nahimana et al.* AC (Chapter 3, n 167) § 839.

to prevent RTLM broadcasts of criminal discourse without having to fear for his own life, or for the lives of his family because he occupied an important position within the interim Government.[175]

277. The Appeals Chamber, further, held that media can cause genocide and constitute persecution, that media leaders can be held responsible for incitement through media or for acts media cause, and that this causal link need not be proven exclusive or essential.[176] The media at issue must 'contribute [...] substantially to the commission of the crime, but [...] it need not be a *sine qua non* condition for its commission.'[177]

3.3.2 Case Vignette: Zyklon B

278. The *Zyklon B* trial was heard in Hamburg before the British Military Court in 1946 and is relevant for our discussion of liability for corporate complicity in mass atrocities. It focused on the complicity of German industrialists in the murder of civilians by means of poison gas. The firm's owner, Bruno Tesch, his subordinate, Karl Weinbacher, and gassing technician Joachim Drosihn were among the accused. Tesch and Weinbacher defended themselves by stressing that they were not on the ground at the extermination camps.[178] The prosecution countered that the fact that they were not physically present was immaterial, claiming that it was illegal to intentionally supply 'a commodity to a branch of the State which was using that commodity for the mass extermination' of civilians.[179] The prosecution utilised testimony from the company's employees' to show that the defendants knew Zyklon B was being used for the 'extermination of human beings', but nevertheless allowed its supply to grow in 'ever-increasing quantities.'[180]

279. While the gassing technician Drosihn was ultimately acquitted, the more senior Tesch and Weinbacher were found guilty.[181] Tesch and Weinbacher, in their roles of senior executives, were found to have materially assisted the crime by guaranteeing a supply of Zyklon B gas, even though they knew it was

175 *Nahimana et al.* AC (Chapter 3, n 167) § 852.
176 MacKinnon (Chapter 3, n 159) 97.
177 *Nahimana et al.* AC (Chapter 3, n 167) §§ 660.
178 *United Kingdom v Tesch* (*The Zyklon B Case*) Case No 9, 1 Law Reports of Trials of War Criminals 93 (British Military Court Hamburg 01–08 March 1946) 97; See also Neema Hakim, 'How Social Media Companies Could Be Complicit in Incitement to Genocide' (2020) 21:1 Chicago Journal of International Law 83, 105.
179 *Zyklon B* (Chapter 3, n 178) 94; Hakim (Chapter 3, n 178) 105.
180 *Zyklon B* (Chapter 3, n 178) 94; Hakim (Chapter 3, n 178) 105.
181 *Zyklon B* (Chapter 3, n 178) 102; Hakim (Chapter 3, n 178) 105.

being deployed to commit mass murder.[182] On the other hand, Drosihn, a technician who lacked decision-making power and distribution control over the gas, was exonerated of any responsibility.[183]

3.3.3 Case Vignette: Charles Taylor

280. Former Liberian President Charles Taylor was convicted of assisting and abetting the atrocity crimes carried out by the Armed Forces Revolutionary Council (AFRC) and Revolutionary United Front (RUF), specifically by giving them material backing and weaponry.[184] The SCSL Appeals Chamber emphasised in its ruling that each case must be considered individually to determine whether an accused person's actions and behaviour had a substantial impact on the commission of the crime.[185] It emphasised that it was impossible to determine whether the behaviour had a substantial impact on the crime's commission up front; rather, that determination had to be reached in the context of all available information. It stated that 'aiding and abetting'-behaviour could involve the supply of weapons and ammunition, financial assistance, the execution of a hate-mongering media campaign, or even the actions of a dentist, accountant, or an architect performing their professional duties.[186]

281. Taylor's conviction was upheld by the Appeals Chamber on the grounds that he had given the RUF/AFRC material support, including weapons, while being aware of the crimes the group was committing and its operational strategy, which was to use a campaign of crimes against the civilian population to further its political and military objectives.[187]

3.4 Liability of Social Media Executives

282. There are two main comparisons which can be drawn when considering the above cases and applying precedents to the potential criminal liability of social media executives. One comparison draws upon the *Fritzsche* and *Nahimana* case to compare social media companies to news outlets and the

182 *Zyklon B* (Chapter 3, n 178) 101–102; Hakim (Chapter 3, n 178) 105.
183 *Zyklon B* (Chapter 3, n 178) 102; Hakim (Chapter 3, n 178).
184 *Prosecutor v Charles Taylor* (Judgement) SCSL-03-01-A [2013] § 362; See also Shannon Raj Singh, 'Move Fast and Break Societies: The Weaponisation of Social Media and Options for Accountability under International Criminal Law' (2019) 8 Chicago International Law Journal 331, 335.
185 *Taylor* AC (Chapter 3, n 184) § 363.
186 *Taylor* AC (Chapter 3, n 184) § 363.
187 *Taylor* AC (Chapter 3, n 184) § 363.

other argues that they are more similar to weapons suppliers, therefore relying on the *Zyklon B* and *Taylor* cases.

3.4.1 Social Media as a News Outlet

283. The comparison of a social media company to an established news outlet, such as Fritzsche's radio or Nahimana's RTLM, has a logical basis both because of the way in which these platforms or outlets directly amplify and spread messaging and due to their similar structure. In Germany and Rwanda mass media outlets played a critical role in inciting violence against certain communities and disseminating degrading speech. Similarly, in Myanmar, the United Nations Fact Finding Mission found that the social media platform Facebook had a 'determining role' in the conflict. Using algorithms designed to prioritise engagement, Facebook amplified posts that dehumanised Rohingya Muslims, spreading these messages to millions.[188]

284. Social media companies are not only comparable to radio and TV in their reach and their ability to inform huge parts of society, they are also similar in structure. Facebook, for instance, has striking structural similarities with RTLM. The radio had, similar to Facebook today, a steering committee that functioned as its board of directors, and it had shareholders.[189] Further, it can be argued that managers and the Chief Executive Officer (CEO) at Facebook have, as the Appeals Chamber in *Nahimana* required 'the power to shape the editorial direction' by having the power to take down posts at their discretion and shape their algorithms in any way they want.[190] This would of course warrant a case-by-case analysis of the actual influence corporate executives have on day-to-day business. They might not have control over the production of content by users, but they have control over the policies that govern what is permitted on their platforms and over the algorithm that decides what content is elevated on their platforms, and what content is suppressed.

285. As mentioned above, the Appeals Chamber in *Nahimana* rejected the notion that the defendants could be convicted merely as owners, publishers, or founders of RTLM.[191] However, under the superior responsibility liability (that can be found in the Rome Statute's Article 28 Section 2) there might be a reasonable argument that CEOs of social media companies have effective

188 Amnesty International, 'The Social Atrocity, Meta and the Right to Remedy for the Rohingya' (2022) 45.
189 *Nahimana et al.* TC (Introduction, n 2) § 566.
190 *Nahimana et al.* TC (Introduction, n 2) § 566.
191 *Nahimana et al.* AC (Chapter 3, n 167) § 594.

control (especially over the algorithm which their own software engineers design and implement), knew what was going on, and did not take the necessary and reasonable measures to prevent such acts.

286. On the other hand, social media companies such as Facebook differ from historical media in terms of the scope and velocity of their propagation of messages. Facebook's algorithm-driven business model enables user-generated material to instantaneously reach viewers throughout the world, whereas RTLM's broadcasts depended on centralised radio transmission. This decentralised structure complicates liability issues by giving offenders more power and weakening the platform's direct oversight. Further, there is another important distinction to be made. The Trial Chamber in Nahimana relied on, although it cannot be established to what extent, the fact that he played an active role in determining the content of RTLM broadcasts, writing editorials and giving journalists texts to read.[192] This suggests that not only the dissemination but the creation of incendiary content might be one factor in determining criminal liability of social media CEOs.[193] However, while the Appeals Chamber in *Nahimana* sustained the Trial Chamber's verdict on direct and public incitement to genocide, they did so under the superior responsibility standard, not mentioning his own speeches, which might indicate that criminal liability for managers can be established even if they do not generate inciting speech themselves.

287. It remains, after all, questionable if the superior responsibility standard applied in *Nahimana* could be useful for social media cases. Effective control could be constituted by the power to delete content and restrict posting and by influencing how the algorithm pushes certain content. This would be reminiscent of how the AC in *Nahimana* established his *de facto* control by arguing he had the power to stop broadcasting in early July 1994.[194] One must remain cautious because in *Nahimana* the defendant's subordinates disseminated inciting speech, while in Facebook's case it was spread through third parties on Facebook's platform. Even though *de facto* control could be sufficient to constitute superior responsibility, the jurisprudence of the ad hoc trials and the ICC lack a single case in which they established superior responsibility outside the military context or a *de jure* superior-subordinate relationship.[195] Users who post on social media platforms cannot be described as subordinates of social media CEOs, but social media company staff, including the senior executives

192 *Nahimana et al.* TC (Introduction, n 2) § 566.
193 Singh (Chapter 3, n 184) 333.
194 See Case Vignette: Nahimana.
195 Wilson and Gillett (Key Guidelines, n 1) § 209.

responsible for devising and implementing content regulation policies and the engineers who write the code for the attention economy algorithms, may be.

288. At the ICC, the *mens rea* element of superior responsibility requires that the civilian superior either knew or 'consciously disregarded information' that indicated the subordinate was acting criminally.[196] This is a weaker standard than that adopted by the United States Supreme Court in the corporate criminal context, which is similar to the test for military officers. It creates an incentive for the superior to search out information (compare 'highest standards of foresight and vigilance' with 'had reason to know'). In contrast, the ICC civilian standard does not create any such incentive because liability attaches if the superior consciously disregards available information.[197] Measures to prevent subordinates from committing crimes in the context of social media could consist in the strict implementation of internal guidelines, alert mechanisms, and local offices in at-risk areas.[198]

3.4.2 Social Media as a Weapon Supplier

Comparisons between social media companies and traditional news outlets like radio and TV might, however, be insufficient to capture their full contribution to mass atrocities. Another alternative is to treat social media companies as analogous to weapon suppliers insofar as they may provide the necessary instruments to facilitate mass atrocities. International criminal tribunals have adjudicated cases involving weapons suppliers and suppliers have been found guilty of aiding and abetting mass crimes.[199] The *Taylor* case at the SCSL and the *Zyklon B* case at the British Military Court serve as precedents.[200]

289. Social media platforms obviously cannot be defined as weapons in the traditional sense.[201] Nonetheless, there are some analogous characteristics, especially when algorithms amplify hateful content and incite violence.[202] As one commentator notes, '[m]any social media platforms today are not passive streams of content submitted by users, but specifically target certain content towards certain demographics in order to procure greater user engagement with the platform. A number of reports and studies have detailed how social media platforms serve to drive extremism through algorithms developed to

196 Slye (Chapter 3, n 138) 966–367.
197 Slye (Chapter 3, n 138) 967.
198 Singh (Chapter 3, n 184) 340.
199 See Case Vignette: Charles Taylor and Case Vignette: Zyklon B.
200 Singh (Chapter 3, n 184) 334–355.
201 Singh (Chapter 3, n 184) 336.
202 Singh (Chapter 3, n 184) 337.

target primal negative human emotions.'[203] Some even argue that 'if you build your algorithms in such a way that it promotes hateful content and inciteful content to the top of someone's news feed, you're doing more than being just a neutral hosting platform.'[204] The intentional design of algorithms that prioritise divisive content for engagement further transforms these platforms into figurative weapons aimed at vulnerable populations.[205] For instance, the Human Rights Council's Fact-Finding Mission in Myanmar identified Facebook as having a 'significant' role in giving hate speech, which incited violence and ethnic cleansing against the Rohingya.[206] The use of Facebook by Myanmar's military to incite and pursue atrocities that have been characterised as potentially genocidal by the United Nations IIMM is not in dispute, and Facebook has since publicly accepted responsibility for its central role in these atrocities.[207]

290. Officers of the Tatmadaw (the official name of Myanmar's military forces) spread propaganda framing the Rohingya as 'dogs,' 'maggots,' and existential threats, contributing to widespread violence, displacement, and acts that have been characterised as genocide. Despite warnings about the potential misuse of the platform, Facebook and its managers failed to take action to mitigate the harm by e.g. removing the content.[208] This pattern is arguably similar to weapon suppliers prioritising profit over responsibility and accountability.[209]

291. Moreover, both weapons suppliers and social media companies are usually far removed from the crime scene.[210] As a matter of criminal law, the accused need not be physically present to hold them accountable for aiding

---

203 Singh (Chapter 3, n 184) 337.

204 Janet H Anderson, 'Liking Genocide on Facebook: Myanmar Rohingya' (*Justice Info* 04 February 2019) <https://www.justiceinfo.net/en/40164-liking-genocide-on-facebook-myanmar-rohingya.html> accessed 27 June 2026; See also Emma Irving, '"The Role of Social Media is Significant": Facebook and the Fact Finding Mission on Myanmar' (*OpinioJuris* 07 September 2018) <http://opiniojuris.org/2018/09/07/the-role-of-social-media-is-significant-facebook-and-the-fact-finding-mission-on-myanmar/> accessed 24 June 2026.

205 Singh (Chapter 3, n 184) 337.

206 Singh (Chapter 3, n 184) 340.

207 Paul Mozur, 'A Genocide Incited on Facebook, with Posts from Myanmar's Military' (*The New York Times* 15 October 2018) <https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html> accessed 7 June 2026; Alexandra Stevenson, 'Facebook Admits It Was Used to Incite Violence in Myanmar' (The New York Times 06 November 2018) <https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html> accessed 07 June 2026.

208 Mozur (Chapter 3, n 207).

209 Mozur (Chapter 3, n 207).

210 Hakim (Chapter 3, n 178) 114.

and abetting.[211] As the judges in the *Taylor* ruling noted, aiding and abetting can occur 'at a time and place removed from the actual crime.'[212] Whereas the ICC Appeal Chamber Judgement in Bemba suggested that distance from the location of the crime was a factor against liability, it was not explicit on the parameters of this putative criteria and did not provide a jurisprudential basis for it. Consequently, the *Bemba* Majority Appeal Opinion runs against the well-established jurisprudence of several tribunals across many cases in this respect. The prevailing view is more compelling given that social media's global reach allows harmful content to influence people thousands of miles away.[213] On a related note, in relation to the liability of business leaders (and presumably all aiders and abetters) some authors advocate aligning the interpretation of Article 25(3)(c) Rome Statute with the jurisprudence of the ad hoc tribunals.[214] That would match the recommendation set out above that the ICC should adhere to the ad hoc tribunals' approach to aiding and abetting as far as possible.

292. Because of the lack of clarity regarding the *actus reus* standard under aiding and abetting, it will likely be difficult to establish what constitutes a 'substantial contribution.'[215] While Facebook's role in Myanmar might be considered sufficiently substantial, other cases might be more contentious. The low causality standard under ICC jurisprudence, however, suggests that the *actus reus* is likely not the highest hurdle to take in this scenario.[216]

293. The *mens rea* requirement might be more arduous to prove. The CEO and managers at Facebook had knowledge about the crimes committed and this might align with the intent requirement for adding and abetting, at least under ad hoc tribunal jurisprudence.[217] While the flagging of inciting language does not necessarily mean that Facebook and its CEO had knowledge about the use of its platform by the Tatmadaw, the numerous warnings by civil society organisations and a meeting at Palo Alto prove that managers at Facebook were aware of how the platform was used and consciously disregarded the information.[218] Both, the *Taylor* and the *Zyklon B* court noted that it sufficed

211 Wilson and Gillett (Key Guidelines, n 1) § 187.
212 *Taylor* AC (Chapter 3, n 184) § 370.
213 Singh (Chapter 3, n 184) 337.
214 Caspar Plomp, 'Aiding and Abetting: The Responsibility of Business Leaders under the Rome Statute of the International Criminal Court' (2014) 30:79 Utrecht Journal of International and European Law 4, 15; Wilson and Gillett (Key Guidelines, n 1) § 196.
215 Plomp (Chapter 3, n 214) 10.
216 Wilson and Gillett (Key Guidelines, n 1) § 83.
217 Wilson and Gillett (Key Guidelines, n 1) § 83.
218 Hakim (Chapter 3, n 178) 114.

that the defendants had knowledge of the atrocities committed with the materials they provided. Under the Rome Statute which demands purpose, it might be harder to convict for aiding and abetting. The Rome Statute, in contrast to the jurisprudence of the ad hoc tribunals, requires proof that the accused acted 'for the purpose of facilitating' a crime.[219] This has been interpreted to mean more than mere knowledge.[220]

294. The ICC version of the mental element for aiding and abetting creates a barrier to prosecution for this form of liability.[221] Applying the purpose standard of the ICC would likely not be met in social media cases because managers and CEOs often do not share the intention of facilitating the crime and even if they did, this would be hard for the Prosecution to prove.[222] The knowledge standard of the ad hoc tribunals (and most domestic law of complicity), however, would be an attainable standard for the prosecution of social media CEOs.

295. If adopted, the analogy to weapon suppliers could fulfil the prevention objective of international criminal law. International criminal law jurisprudence has been somewhat successful in deterring arms dealers from supplying weapons to ongoing armed conflicts because they incorporate the prospect of criminal prosecution into their cost-benefit analyses.[223] Extending social media companies' liability to aiding and abetting could have the same effect, incentivising social media CEOs and managers to take proactive measures against their platforms' misuse.[224]

296. Legal scholar Hakim stratifies layers of liability for different positions held within the company. He finds possible liabilities (depending on the knowledge) for positions such as CEO and managers but suggests that there might not be liability for content managers. CEOs, while fulfilling the *actus reus* by providing a platform for inciting speech, might not always satisfy the *mens rea* requirement because they are less likely to know the day-to-day details of the company.[225] The case of managers might be somewhat easier for they might satisfy the *actus reus* by being involved in everyday business and are often aware of what is happening in their area of management.[226] However, this is probably most true for senior officials rather than mid-level officials and here we recall the *Zyklon B* cases in which the high-ranked company officials were

219 Rome Statute (Key Guidelines, n 2) art 25 (3)(c).
220 Plomp (Chapter 3, n 214) 15.
221 Wilson and Gillett (Key Guidelines, n 1) § 195.
222 Wilson and Gillett (Key Guidelines, n 1) § 195.
223 Singh (Chapter 3, n 184) 337.
224 Singh (Chapter 3, n 184) 337.
225 Hakim (Chapter 3, n 178) 110.
226 Hakim (Chapter 3, n 178) 114.

convicted.[227] Similarly, the lack of material decision-making authority makes criminal liability for content moderators unlikely, just as the lower-level functionary Drosihn was acquitted in the *Zyklon B* cases.[228]

3.4.3 Other Avenues

297. Corporate criminal liability offers a promising avenue for holding platforms accountable, as can be seen in recent cases against industrial and corporate actors. Additionally, regulatory frameworks enshrining proactive content moderation and transparency could limit the misuse of social media. At the UN, there has been a proposal for a Code of Conduct for Information Integrity on Digital Platforms.[229] The Code of Conduct would guide states, private social media companies, and other stakeholders in their efforts to make the digital space safer while protecting freedom of speech.[230] This would not only hold states accountable for the questionable dimensions of their information campaigns on social media but could more clearly define the obligations that private social media companies themselves face.[231] A Code of Conduct should enhance transparency and inculcate a context-specific approach to content moderation.[232] Further, civil remedies, such as class action lawsuits, anti-trust litigation, and even an International Court of Torts may complement these strategies by providing redress for victims.

298. Comparisons of social media and traditional media illustrate both the inculpatory and exculpatory implications for companies in the digital era. There are valid analogies between social media companies and traditional media outlets such as radio and TV as both easily reach vast parts of society. However, the jurisprudence does not support a simple and direct application to social media executives of the superior responsibility standard under Article 28

227 Hakim (Chapter 3, n 178) 115.

228 Hakim (Chapter 3, n 178) 115.

229 UN, 'SG's Policy Brief on Information Integrity on Digital Platforms' (*UN* 2024) <https://www.un.org/en/civil-society/information-integrity-digital-platforms> accessed 09 January 2025.

230 Article 19, 'UN: Comments on Code of Conduct for Information Integrity on Digital Platforms – ARTICLE 19' (*Article 19* 20 February 2024) <https://www.article19.org/resources/un-comments-on-code-of-conduct-for-information-integrity-on-digital-platforms/> accessed 09 January 2025.

231 UN Comments Article 19 (Chapter 3, n 230).

232 Richard A Wilson and Molly K Land, 'Hate Speech on Social Media: Content Moderation in Context' (2021) 52 Connecticut Law Review 1029; Richard A Wilson and Jordan Kiper, 'Incitement in an Era of Populism: Updating Brandenburg after Charlottesville' (2020) 5 University of Pennsylvania Journal of Law and Public Affairs 56.

of the Rome Statute. In particular, users posting on social media cannot constitute subordinates, but platform designers and executives who perform oversight might in future litigation. *Nahimana* is instructive when it comes to the potential criminal liability of social media CEOs because it provides guidance on the criteria considered by courts when establishing responsibility.

299. Therefore, the analogy of social media companies to weapon suppliers and their consideration under a mode of liability of aiding and abetting might be more appropriate. Cases such as *Taylor* and *Zyklon B* are valuable precedents. Social media can have the destructive power of a weapon because algorithms are often programmed to promote highly engaging content (regardless of its inciting content), which may favour inflammatory and hateful messaging to attract engagement. Further, the physical absence from the scene of the crime makes social media companies comparable to weapon suppliers. Aiding and abetting might be useful also because of its low causality standard. However, the heightened *mens rea* standard ("purpose") at the ICC might prove to be a barrier to prosecution. Arguably, without a change in jurisprudence or an amendment of Article 25(3)(c) of the Rome Statute, prosecuting social media CEOs under this mode of liability presents difficulties. Further, there might be differences in potential criminal liability whether one prosecutes a CEO, a manager or a content moderator.

4 Establishing Jurisdiction over Speech Crimes Disseminated through the Internet, including on Social Media Platforms

300. As recognised in the Policy on Cyber-Enabled Crimes under the Rome Statute published by the Office of the Prosecutor in December 2025,[233] asserting territorial jurisdiction over speech crimes disseminated through the internet may obstruct prosecutions of alleged speech crimes.[234] In light of the demonstrable role played by social media platforms and other digital spaces in ongoing conflict situations, this merits further examination.

301. The ICC's jurisdiction is established through the Rome Statute when a situation fulfils certain statutory preconditions: temporal jurisdiction, territorial or personal jurisdiction, and subject-matter jurisdiction.[235] While this

233 International Criminal Court Office of the Prosecutor, 'Policy on Cyber-Enabled Crimes under the Rome Statute' (ICC OTP December 2025) § 3. Hereafter also referred to simply as the Policy.

234 Wilson and Gillett (Key Guidelines, n 1) §§ 328–330.

235 Rome Statute (Key Guidelines, n 2) art 5 (11–12).

applies to any case investigated by the Office of the Prosecutor (OTP), including when the alleged crimes under investigation are enabled or committed by cyber means, the instrumental role of digital technologies and networks in conflict settings generates several unique obstacles. In March 2025, the OTP launched a public consultation, recognising that several of the crimes addressed within the Statute 'may be committed or facilitated by cyber means, and that the Court's jurisdictional framework can apply to them'.[236] The resulting Policy on Cyber-Enabled Crimes under the Rome Statute reiterates that 'any case investigated by the Office – including with regard to cyber-enabled crimes – must form part of a situation fulfilling the statutory preconditions to the exercise of the Court's jurisdiction: temporal jurisdiction; either territorial or personal jurisdiction; and subject-matter jurisdiction'.[237]

302. The Policy considers core questions, such as the application of the ICC's jurisdictional framework to cyber-enabled crimes, based on the long-standing international law principles of territorial and personality jurisdiction and how the crimes under the Rome Statute may be either directly committed or facilitated through cyber means. It also highlights a number of practical considerations, including the need for technical expertise, access to evidence, and cooperation with national authorities and the private sector. The development of the Policy on Cyber-Enabled Crimes under the Rome Statute marks a vital step towards strengthening accountability and improving the mechanisms for investigating and prosecuting international crimes enabled or facilitated through cyber conduct.

303. While the Policy addresses briefly the topic of territorial jurisdiction over cyber-enabled crimes, this requires further clarification. For example, the Policy touches upon situations where online conduct may take place simultaneously across the territorial boundaries of multiple States at the same time. Regarding such situations, it states 'that the use of cloud computing technology to commit or facilitate crimes under the Rome Statute may pose one such example' where the territoriality principle could equally apply.[238] It also notes that there remains disagreement among both States and experts about the application of the territoriality principle in situations 'where there is only a minimal connection with cyber infrastructure on a State's territory', highlighting that 'the Office does not presently anticipate that it would regard the mere

236 International Criminal Court Office of the Prosecutor, 'Draft Policy on Cyber-Enabled Crimes under the Rome Statute' (ICC OTP March 2025) § 3b.

237 ICC OTP Cyber Policy (Chapter 3, n 236) § 38.

238 ICC OTP Cyber Policy (Chapter 3, n 236) § 44.

transit of data through a State Party's territory as a sufficient basis to assert the Court's territorial jurisdiction'.[239]

304. The actual threshold triggering jurisdiction remains unclear and should be clarified in accordance with the standards set forth in the Tallinn Manual on the International Law Applicable to Cyber Warfare (Tallinn Manual 2.0) Rule 9, which considers that '[a] State may exercise territorial jurisdiction over: (a) cyber infrastructure and persons engaged in cyber activities on its territory; (b) cyber activities originating in, or completed on, its territory; or (c) cyber activities having a substantial effect in its territory.'[240]

305. The section below expands on the emerging legal framework governing territorial jurisdiction over crimes committed online raised by the Policy and considers how it applies specifically to the issue of speech crimes enabled, facilitated or committed via digital means – including on social media platforms. Given the global reach of the medium, speech disseminated online over social media that may amount to a speech crime (e.g. direct and public incitement to genocide) transcends the territory of any one State. For example, while the author of an inciting social media post is based in one State, the platform's host server may be located in another, and the potential harmful consequences of said post may have a deleterious impact on a conflict occurring yet elsewhere. Given the transboundary nature of online speech acts, it should be considered whether – and if so, on what basis – the ICC can establish territorial jurisdiction for investigating and prosecuting speech crimes occurring in the cyber realm and what challenges may need to be overcome to do so.

306. With regard to the territoriality principle, it is necessary to define both the scope of the conduct of relevance and to determine where it has taken place.[241] Lessons can be drawn from domestic courts and their responses to crimes occurring in online spaces. Germany and France, for instance, have both developed and accepted an audience-based approach through the 'effects doctrine'.[242] Such an approach establishes jurisdiction on the basis of the target audience location and the spread of the harmful effect of an offence, rather than the physical location where it is actually committed.

307. Turning to international criminal law, under the limited and specific nature of the ICC's jurisdictional framework it is questionable whether such

239 ICC OTP Cyber Policy (Chapter 3, n 236) § 45.

240 International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence, Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Michael Schmitt ed, Cambridge University Press 2017) 55.

241 ICC OTP Cyber Policy (Chapter 3, n 236) § 42.

242 See the discussion which follows in ICC OTP Cyber Policy (Chapter 3, n 236) § 42–48.

an approach could apply. However, at the ICC the ruling of Pre-Trial Chamber I in the Myanmar Jurisdiction Decision offers some insight into how the Court might establish jurisdiction in cases with transboundary offences. This Decision suggests that the Court has the ability to exercise jurisdiction on a territorial basis because an element or part of an alleged crime took place on the territory of a State party – a potential way forward which nevertheless remains particularly complex for certain modalities of crime. For example, in the case of incitement, different countries might be involved (country of author, website, host-server, …) and the formulation as a mode of liability might make it necessary to establish the facticity of actual genocidal acts. Verifying the material aspects of the crime may also require establishing a direct causal connection (using the "significant contribution" standard) between the conduct and the result – particularly where they are of far proximity or where the conduct takes place online – and the gravity of the harm.[243]

308. This sub-section first explores briefly the approach taken by domestic Courts to cyber-enabled speech crimes. It then discusses the implications of ICC Pre-Trial Chamber I in the Myanmar Decision and finally considers the obstacles to ICC jurisdiction over core international crimes committed on social media platforms and how these might be overcome.

5 The Approach Taken by Domestic Courts to Criminal Activity Online

309. Although the Council of Europe and the EU have addressed the punishment of crimes committed through the internet, and some countries have developed specific legislation in this area, there is no general international consensus among domestic legal systems on the modalities of prosecuting online criminality.[244] The following section considers key cases from Germany and France concerning criminal speech acts or displays online which may be instructive on how some domestic legal systems deal with the issue. These cases indicate that when seeking to establish jurisdiction over cyber-enabled crimes in this area, domestic courts have adopted an approach which tests whether one (or the most) part(s) of the crime occurred on the State(s) territory.

310. The 1999 *Töben* case, which was ultimately brought before the German Federal Court of Justice, concerned a series of blogposts made by Australian

243 ICC OTP Cyber Policy (Chapter 3, n 236) footnote 45, citing Tallinn Manual 2.0 (Chapter 3, n 240) 55.

244 Xin Wang, 'Global (Re-)Framing of Cybercrime: An Emerging Common Interest in Flux of Competing Normative Powers?' (2024) 38:2 Leiden Journal of International Law 235, 260.

national Frederick Töben using a host server in Australia in which he effectively denied the Holocaust.[245] Although the website was published in the English language, it was accessed in Germany by German neo-Nazis.[246] Australia does not criminalise Holocaust denial, however, such speech can amount to criminal conduct in Germany where due to the specific political and historical context it is viewed as a form of incitement if it results in disturbing the public peace.[247]

311. In *Töben* the German Federal Court of Justice construed territorial jurisdiction on the basis that the harmful effects of the posts 'could be construed as having occurred in Germany'.[248] The Court further argued that although the statements were in English, they were addressed to the German public and accessible to German users.[249] On this basis, the Court held that its assertion of jurisdiction was in accordance with public international law – noting a substantial connection between the State exercising jurisdiction and the offence at hand 'because the harmful effects, one of the elements of the offence, took place in Germany'.[250]

312. Similarly, French courts asserted jurisdiction in the so-called *Yahoo!* case arguing that the act in question had materialised equally in French territory since the Yahoo! websites were accessible from France.[251] The case involved the display and selling of Nazi memorabilia on the platform Yahoo!. The courts held that the visualisation of the objects by itself caused an injury and, thus, the injury occurred on French territory.[252] The Court of Appeals, while upholding the Lower Court's reasoning, added that a more restrictive view of its jurisdiction would likely promote a safe haven for online hate speech, thus, advancing a judicial policy argument.[253]

---

245 Wilson and Gillett (Key Guidelines, n 1) § 328; Michail Vagias, 'The Territorial Jurisdiction of the ICC for Core Crimes Committed through the Internet' (2016) 21 Journal of Conflict and Security Law 523, 531; *Re Töben* (Decision of the District Court of Mannheim) Urt. v 10.11.1999-5 KLs 503 Js 9551/99; Decision of the Federal Court of Justice (BGH), Urt. v 12.12.2000-1 StR 184/00, reported in 54(8) NJW (2001) 624–628.

246 Joachim Vogel, 'Towards a Global Convention against Cybercrime' (2007) First World Conference of Penal Law.

247 StGB (Chapter 2, n 228) s130(3)-(4).

248 *Decision of the Federal Court of Justice* (BGH) (Chapter 3, n 245) 624–628.

249 Vagias (Chapter 3, n 245) 532; *Re Töben* (Chapter 3, n 245).

250 Vagias (Chapter 3, n 245) 532; *Re Töben* (Chapter 3, n 245).

251 *Licra v Yahoo!* (Tribunal de grande instance [TGI] [Ordinary Court of Original Jurisdiction]) [2000] RG:00/0538 (Fr.).

252 Matthias Reimann, 'Introduction: The Yahoo! Case and Conflict of Laws in the Cyberage' (2003) 24 Michigan Journal of International Law 663, 663.

253 Vagias (Chapter 3, n 245) 533.

313. While these decisions have received some criticism,[254] similar litigations in different countries seem to follow the same pattern.[255] However, in *R v. Sheppard and Whittle* the English Court of Appeal applied different reasoning albeit with the same outcome.[256] The Court applied the 'substantial measure' test, requiring that most acts of the crime must have had to occur on UK territory.[257] The fact that the server the material was hosted on was not on UK territory was not enough to vitiate the court's jurisdiction.[258]

314. These cases indicate an approach developing across several different domestic jurisdictions which establishes jurisdiction on the basis that an element of the crime occurred on the State's territory – including the effect of the crime. It is of note that this focuses on the act of 'display' through the use of an online system – rather than the actual act of typing or posting – and that the decision in the *Töben* case specifically emphasises the consequences and harmful effect of the crime. The so-called 'effects doctrine' as applied in *Töben* has emerged as a potentially substantial extension of the territoriality principle in State practice, particularly in anti-Trust cases.[259] Given the transboundary nature of cyber-enabled speech crimes, this approach might be particularly useful for establishing jurisdiction in digital cases.

315. Although this development was acknowledged in the Policy, the ICC Office of the Prosecutor has so far taken the limiting view that 'the territoriality principle as articulated in the Court's jurisprudence and applied to the cyber context [...] appears to be sufficiently flexible to cover the great majority of situations involving cyber-enabled crimes likely to be of relevance to the Office. Therefore, it does not consider it necessary at this time, to take a position on the effects doctrine and its potential applicability to cyber-enabled crimes within the subject-matter jurisdiction of the Court'.[260]

---

254 Vagias (Chapter 3, n 245) 533.

255 See among others, *Valve Corporation v Australian Competition and Consumer Commission* [2017] FCAFC 224–258 FCR 190; *Google Spain SL, Google Inc. v Agencia Española de Protección de Datos* (*AEPD*) *and Mario Costeja González* 62012CJ0131 (Court of Justice of the European Union (CJEU) 13 May 2014).

256 Julia Hörnle, 'The Internet and Criminal Jurisdiction' (*Society for Computers and Law* 18 March 2010) <https://www.scl.org/1786-the-internet-and-criminal-jurisdiction/> accessed 26 June 2026.

257 Hörnle (Chapter 3, n 256).

258 Hörnle (Chapter 3, n 256).

259 This also holds true for other areas – see anti-trust context as discussed in Najeeb Samie, 'The Doctrine of 'Effects' and the Extraterritorial Application of Antitrust Laws' (1982) 14:1 University of Miami Inter-American Law Review 23.

260 ICC OTP Cyber Policy (Chapter 3, n 236) § 43.

6 At the ICC

316. Developments in State practice should be considered alongside the precedent set by Pre-Trial Chamber (PTC) I and III of the ICC regarding the situation in Myanmar in 2018 and 2019, respectively. Although the decisions were specifically concerned with physical acts relating to the alleged deportation of members of the Rohingya people from Myanmar, a non-State party to the Rome Statute, to Bangladesh, a State Party they nevertheless offer potential insights that may be relevant to asserting jurisdiction over online crimes in the future.

317. In the 2018 Myanmar Decision, PCT I deliberated on the issue of whether the Court could exercise jurisdiction based on the territoriality principle when the crime in question took place across the territories of multiples States, including when one (Myanmar) was not a Party to the Rome Statute. PTC I ultimately held that the Court could establish territorial jurisdiction because it was sufficient for an element or part of the alleged crime to have taken place on the territory of a Party (Bangladesh).[261] This was followed in 2019 by the decision of PTC III concerning the same situation, which authorised the Prosecutor's request to initiate an investigation into the alleged crime which took place at least in part in the territory of Bangladesh and which mirrored the reasoning of PTC I.[262]

318. The reasoning behind the Myanmar Decision is grounded in Article 12(2)(a) of the Rome Statute, which provides that the Court may exercise jurisdiction if 'the State on the territory of which the conduct in question occurred' is Party to the Rome Statute.[263] By using an 'object and purpose' approach – meaning that at least one legal element or part of the crime has to occur on the territory of the State Party – to interpreting Article 12(2)(a) PTC I takes an expansive interpretation.[264] It therefore holds that given the transnational character of the crime, any other interpretation would undermine the Court's ability to prosecute perpetrators, as all countries in which a part of the crime occurred would have the be State Parties.[265]

---

261 Decision on the Prosecution's Request for a Ruling on Jurisdiction under Article 19(3) of the Statute, Request under Regulation 46(3) of the Regulations of the Court (Decision, Pre-Trial Chamber I) ICC-RoC46(3)-01/18-37 (6 September 2018).

262 Decision Pursuant to Article 15 of the Rome Statute on the Authorisation of an Investigation into the Situation in the People's Republic of Bangladesh/Republic of the Union of Myanmar (Decision, Pre-Trial Chamber III) ICC-01/19-27 (14 November 2019) §§ 123–134.

263 Rome Statute (Key Guidelines, n 2) art 12(2)(a).

264 Caroline Sweeney, 'Accountability for Syria: Is the International Criminal Court Now a Realistic Option?' (2019) 17:5 Journal of International Criminal Justice 1083, 1085, 1093.

265 PTC I Article 19(3) Jurisdiction Decision (Chapter 3, n 261) § 74.

319. Indeed, PTC I clearly emphasised that the occurrence of the consequence of the crime on State Party territory suffices to establishes territorial jurisdiction, arguing that the exercise of jurisdiction can encompass 'acts of deportation initiated in a State not Party to the Statute (through expulsion or other coercive acts) and completed in a State Party to the Statute (by virtue of victims crossing the border to a State)'.[266] Significantly, PTC I held in its decision that the reasoning it applied in the Myanmar Jurisdiction Decision applies by analogy to other crimes with a potential cross-border dimension – setting a significant precedent.[267]

320. In authorising the Prosecutor to investigate the situation in Bangladesh and Myanmar, PTC III also considered the meaning of the term "conduct" under Article 12(2)(a) of the Rome Statute, finding that it included the consequences of conduct (or perlocutions, in the case of speech acts). In line with the decision by PTC I, PTC III held that 'at least part of the conduct [...] must take place in the territory of a State Party'.[268] Turning to customary international law regarding this principle, PTC III cited that 'States are free to assert territorial criminal jurisdiction, even if part of the criminal conduct takes place outside its territory, as long as there is a link with their territory'.[269] In line with the reasoning also applied by PTC I, PTC III asserted that 'it would be wrong to conclude that States intend to limit the Court's territorial jurisdiction to crimes occurring exclusively in the territory of one or more States Parties'. Moreover, reading Article 12(2)(a) of the Statute in this manner would go against the principle of good faith (including effective interpretation).[270] That jurisprudence has since been applied in the context of the Lithuania/Belarus situation. The Office of the Prosecutor concluded that there is a reasonable basis to believe that Rome Statute crimes (such as deportation and persecution) were committed in part in Lithuania, despite the crimes being initiated in Belarus.[271]

321. This interpretation was not without criticism, with Myanmar countering that under the principle of *pacta tertiis nec nocet nex preosunt* (treaties cannot create obligations for third parties without their consent) the ICC should be barred from extending jurisdiction in this manner.[272] The decision to conflate

266 PTC I Article 19(3) Jurisdiction Decision (Chapter 3, n 261) § 73.
267 Sweeney (Chapter 3, n 264) 1093.
268 PTC III Bangladesh/Myanmar Authorisation Decision (Chapter 3, n 262) § 50, 61.
269 PTC III Bangladesh/Myanmar Authorisation Decision (Chapter 3, n 262) § 58.
270 PTC III Bangladesh/Myanmar Authorisation Decision (Chapter 3, n 262) § 59.
271 Office of the Prosecutor, 'The Situation in the Republic of Lithuania / Republic of Belarus: Summary of Preliminary Examination Findings', 12 March 2026.
272 Sweeney (Chapter 3, n 264) 1095; Annex E to Prosecution Notice of Documents for Use in Status Conference (Pre-Trial Chamber) (19 June 2018) ICC-RoC46(3)-O1/18/27/AnxE.

conduct with consequence has also been criticised by some,[273] and was itself subject to a dissenting opinion.

322. The 2019 PTC II Afghanistan Investigation Decision is also of relevance here. PTC II seemingly aligned itself with the position taken by PTC I and III regarding the territorial jurisdiction and the meaning of 'conduct' asserting that 'the Court has jurisdiction if the conduct was either completed in the territory of a State Party or if it was initiated in the territory of a State Party or vice versa'.[274]

323. Other aspects of this decision are worthy of further consideration. The OTP submitted that, in cases where a suspect is not the direct perpetrator of a crime, but rather facilitates it (whether through accessorial conduct or other forms of responsibility) 'a suspect is not required to be physically present in the territory of a State Party when a crime is committed for the Court to be able to exercise jurisdiction over his or her conduct, as long as the crime imputed to the suspect occurred within the confines of such territory'.[275]

324. The OTP further submitted that, 'a suspect's own role or conduct may thus be geographically and temporarily removed from the territory where the crime occurs, as long as the underlying conduct which is imputed to the suspect occurs within the Court's jurisdiction. As such, forms of attribution are not dependent on proximity: a suspect may be wholly outside of a State's territory when, for example, committing a crime through or with another person or persons, or ordering its commission, or otherwise facilitating its commission by aiding and abetting – he or she can still be held criminally responsible under the Statute if all of the material elements of the underlying criminal conduct which is imputed to the suspect occurs within the Court's jurisdiction'.[276] Although these submissions were based on references to case law, the PTC did not directly address these submissions in its decisions on the authorisation of an investigation.

---

273 Caleb H Wheeler, 'Human Rights Enforcement at the Borders: International Criminal Court Jurisdiction over the Rohingya Situation' (2019) 17:3 Journal of International Criminal Justice 609, 627.

274 Decision Pursuant to Article 15 of the Rome Statute on the Authorisation of an Investigation into the Situation in the Islamic Republic of Afghanistan (Decision, Pre-Trial Chamber II) ICC-02/17-33 (12 April 2019) § 50.

275 Public redacted version of 'Request for authorisation of an investigation pursuant to article 15', 20 November 2017 ICC-02/17-7-Conf-Exp, Situation in the Islamic Republic of Afghanistan (ICC-02/17-7-Red), Office of the Prosecutor, 20 November 2017 § 47.

276 Request for authorisation of an investigation pursuant to article 15 (Chapter 3, n 275) § 47.

325. The Prosecutor's request to investigate the situation in Afghanistan was ultimately denied by PTC II, which determined that such an investigation did not serve the interests of justice primarily due to feasibility issues. Nonetheless, this was then amended by the Appeals Chamber in 2020 which authorised an investigation.[277]

326. In any case, the lesson we draw here is that the PTC II provided grounds for jurisdiction over speech crimes committed on social media or other online platforms even if the speaker is geographically removed from the location of physical perpetrators.

7 Social Media Speech Regulation

327. In addition to inciting genocide online, governments engage in a number of online practices that violate human rights.[278] Across the world, governments and state-aligned actors orchestrate online harassment campaigns against racial and religious minority groups and political opponents.[279] Government online propaganda operations have been termed 'digital authoritarianism,'[280] and are characterised by an array of anti-democratic techniques that include internet shutdowns, surveillance,[281] censorship of online speech, disinformation,

277 PTC II Afghanistan Authorisation Decision (Chapter 3, n 274) §§ 33–35, 43–44, 87–96; in 2020 the Appeals Chamber amended PTC II's decision and authorised an investigation, Judgement on the appeal against the decision on the authorisation of an investigation into the situation in the Islamic Republic of Afghanistan (05 March 2020) ICC-02/17-138.

278 Tamar Megiddo, 'Online Activism, Digital Domination, and the Rule of Trolls: Mapping and Theorizing Technological Oppression by Governments' (2020) 58 Columbia Journal of Transnational Law 394.

279 'State actors' denotes government officials or agencies and 'state-aligned actors' denotes individuals whose speech on social media aligns closely with the interests of the government or military.

280 See Adrian Shahbaz, 'The Rise of Digital Authoritarianism' (*Freedom House* 2018); Erol Yaybroke and Samuel Brannen, 'Promote and Build: A Strategic Approach to Digital Authoritarianism' (*Center for Strategic and International Studies* 2020); Ron Deibert defines authoritarianism as 'state constraints on legitimate democratic political participation, rule by emotion and fear, repression of civil society, and the concentration of executive power in the hands of an unaccountable elite' in Deibert (Chapter 3, n 15) 64.

281 David Kaye, 'Surveillance and Human Rights – Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression' (28 May 2019) A/HRC/41/35 § 26.

state-sponsored trolling,[282] and incitement of violence.[283] Governments coordinate online propaganda operations to intimidate and silence critics and to galvanise popular support for a range of restrictive measures that include the criminalisation of political and legal dissent, and in some contexts, disappearances and killings.[284]

328. Given this evidence, it is no longer credible to maintain that online speech does not have the potential to cause substantial harm and injury. There is ample research showing that in certain circumstances hate speech correlates with, and possibly has a causal nexus to, measurable harms. Of course, an empirical consensus on the deleterious effects of inciting speech does not in itself determine the legal or normative questions of hate speech regulation, since one might reasonably support freedom of expression for other reasons, including its value in the process of reporting on mass crimes or democratic decision-making.

329. Governments are no longer the primary regulators of speech.[285] Their regulatory capacity has been far outstripped by some of the largest companies in the world by public stock valuation,[286] which together regulate the speech of several billion active social media users.[287] Facebook/Meta, which moderates the online speech of over three billion active monthly user accounts, is the largest publisher of content in human history.[288]

---

282 'State-sponsored trolling' is the coordination by an official state agency or party in government of automated accounts (bots), paid users, and volunteers to harass opponents. See Carly Nyst and Nick Monaco, 'State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns' (Institute for the Future 2018) 5 defining and analysing trolling.

283 See Muna Abbas and others, 'Invisible Threats: Online Hate Speech against Human Rights Defenders in Guatemala' (American Bar Association Center for Human Rights 2019); Megiddo (Chapter 3, n 278). Wilson, Richard A., 'The Anti-Human Rights Machine: Digital Authoritarianism and the Global Assault on Human Rights.' (2022) 44 Human Rights Quarterly, 704.

284 Abbas and others (Chapter 3, n 283); Megiddo (Chapter 3, n 278).

285 See David Kaye, 'Speech Police: The Global Struggle to Govern the Internet' (2019) Columbia Global Reports – noting that 'a few private companies' control social media.

286 See NASDAQ, 'FB Facebook, Inc. Class a Common Stock' (*NASDAQ* 2022) <https://www.nasdaq.com/market-activity/stocks/meta> accessed 14 January 2020 (listing Facebook's common stock value); NASDAQ, 'GOOG Alphabet Inc. Class C Capital Stock' (*NASDAQ* 2019) <https://www.nasdaq.com/market-activity/stocks/goog> accessed 14 January 2020 (listing Google's capital stock value).

287 Simon Kemp, 'Digital 2019 Q4 Global Digital Statshot' (2019) <https://www.slideshare.net/slideshow/embed_code/key/NSBCWzcfsbGXTg> accessed 30 January 2020 – Slide 6.

288 Facebook, Inc, 'Facebook Reports Third Quarter 2019 Results' (*Facebook* 2019) <https://s21.q4cdn.com/399680738/files/doc_financials/2019/q3/FB-Q3-2019-Earnings-Release.pdf> accessed 1 May 2020; Facebook Statistics 2026: Users, Revenue, Demographics &

330. In 2019, social media companies including Meta, Google, and Twitter (now X), removed seventy-two percent of hate speech on their platforms that violated the European Commission's code of conduct.[289] Facebook/Meta reported in 2017 that it was deleting nearly 288,000 hate speech posts per month,[290] and this tripled to nearly 1 million hate speech posts per month by the third quarter of 2018.[291] In the first quarter of 2020, Meta reported taking action on 9.6 million posts, or 3.2 million per month.[292] The rate of removal can spike even higher during critical incidents, such as when the New Zealand mosque shooter streamed on Facebook Live in 2019.[293] In the twenty-four hours after that tragedy, Meta reported that it prevented the uploading of more than 1.2 million videos of the attack and removed approximately 300,000 additional copies after they were posted.[294]

331. Although the volume of hate speech content removed from Meta platforms has remained high, there is a notable decrease following the implementation of a new content moderation policy. In 2025, Meta made significant changes to its content moderation policies and operations, including the removal of fact-checkers in the US,[295] focusing automated filters only on illegal

---

Advertising Data – Axis Intelligence, https://axis-intelligence.com/facebook-statistics/ accessed 30 June 2026.

289 Elizabeth Schulze, 'EU Says Facebook, Google and Twitter Are Getting Faster at Removing Hate Speech Online' (*CNBC* 4 February 2019) <https://www.cnbc.com/2019/02/04/facebook-google-and-twitter-are-getting-faster-at-removing-hate-speech-online-eu-finds--.html>; The European Commission defines hate speech as 'all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin' in EC, 'Code of Conduct on Countering Illegal Hate Speech Online' (2016).

290 Nadine Strossen, 'Hate: Why we Should Resist It With Free Speech, Not Censorship' 31 (2018) 32.

291 Facebook took action against 2.9 million pieces of content in the third quarter of 2018, a rate of nearly 1 million per month: Mehreen Khan, 'More "Hate Speech" Being Removed from Social Media' (*Financial Times* 04 February 2019) <https://www.ft.com/content/868f9d96-27bc-11e9-a5ab-ff8ef2b976c7> accessed 09 June 2026.

292 Facebook, 'Community Standards Enforcement Report' (*Facebook* 2020) <https://transparency.facebook.com/community-standards-enforcement#hate-speech> accessed 26 May 2020.

293 Nick Perry and Mark Baker, 'New Zealand Mosque Shooter Broadcast Slaughter on Facebook' (*AP News* 16 March 2019) <https://apnews.com/article/shootings-ap-top-news-international-news-arrests-immigration-ce9e1d267af149dab40e3e53912545 30> accessed 10 June 2026.

294 Guy Rosen, 'A Further Update on New Zealand Terrorist Attack' (*Facebook* 20 March 2019) <https://about.fb.com/news/2019/03/technical-update-on-new-zealand/> accessed 10 June 2026.

295 Mark Zuckerberg, (*Facebook* 07 January 2025) 'Fact-checkers replaced with community notes system, similar to that employed on X: Mark Zuckerberg, "It's Time to Get back to

and high-severity violations, increasing the threshold for removal of content, removing certain restrictions on content related to topics such as immigration and gender, and abandoning the system of third-party fact checkers.[296] Meta CEO Mark Zuckerberg justified these changes by referring to a restoration of freedom of speech on Meta's platforms, where 'too much censorship' had occurred.[297] A transparency report published in 2026 by Appeals Centre Europe[298] highlights that in seventy percent of cases where a social media platform allowed content reported as hate speech to remain online, their independent review disagreed with the decision.[299]

332. Social media companies content moderation rules (or 'platform law') employ far-reaching definitions of hate speech. For example, YouTube's policy is to 'remove content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, veteran status.'[300] Indeed, YouTube's policy goes beyond Germany's incitement to hatred laws by prohibiting speech based on message alone and without reference to the potential for real-world impact.[301] X's Hateful Conduct Policy censors tweets that 'promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.'[302] Except for hateful images or symbols in a profile image or header that express hate towards a person, group, or protected category, X's (formerly Twitter) rules allow hate speech as long as it does not incite violence or attack someone on the basis of their membership in a group.[303]

Our Roots around Free Expression"' <https://www.facebook.com/reel/1525382954801931> accessed 10 June 2026.

296 Zuckerberg (Chapter 3, n 295).

297 Zuckerberg (Chapter 3, n 295).

298 Appeals Centre Europe, 'Appeals Centre Europe Transparency Report' (*Appeals Centre Europe* 2026) – Appeals Centre Europe is an independent out-of-court dispute settlement body certified under Article 21 of the European Union's Digital Service Act.

299 'Social Media Platforms Failed to Remove Hate Speech in Most Cases Reviewed in EU, Report Says | MLex | Specialist News and Analysis on Legal Risk and Regulation' (*MLex* 2026) 4 <https://www.mlex.com/mlex/articles/2481558> accessed 26 June 2026.

300 YouTube, 'Hate Speech Policy' <https://support.google.com/youtube/answer/2801939?hl=en> accessed 26 June 2026.

301 See, for example, StGB (Chapter 2, n 228) s130(1) discussing penalties for '[w]hoever, in a manner which is suitable for causing a disturbance of the public peace, incites hatred against a national, racial, religious group or a group defined by their ethnic origin'.

302 Twitter/X, 'Hateful Conduct' (*X* (*formerly Twitter*) *Help Center* April 2023) <https://help.x.com/en/rules-and-policies/hateful-conduct-policy> accessed 26 June 2026.

303 Twitter/X (Chapter 3, n 302).

333. Of the mainstream social media companies, Facebook/Meta's definition of hate speech is the broadest and is implemented by the most aggressive procedures for moderating content. Facebook/Meta defines hate speech as 'a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability.'[304] Facebook/Meta also provides 'some protections for immigration status.'[305] It defines an attack 'as violent or dehumanising speech, statements of inferiority, or calls for exclusion or segregation.'[306]

334. Platform law is enforced by automated content tools and human moderators who assess content that has been reported by users.[307] Harmful online content is classified through both automated moderation tools (algorithms) as well as user flagging.[308] Algorithms identify particular classes of harmful material such as child pornography, extremist content, and hate speech.[309] Additionally, Facebook/Meta and X rely on users to identify harmful content, and each has created a flagging mechanism for users to tag content that is then sent to a human reviewer.[310]

335. Some commentators hailed the oversight board as a 'Supreme Court'[311] that would be 'completely autonomous to Facebook'[312] and 'can overrule Zuckerberg.'[313] However, its authority is quite limited, and its autonomy is circumscribed in important ways. Specifically, there is no provision in the bylaws for the board to review material that remains posted on the platform,

304 Facebook, Inc, 'Policies: Community Standards – Hateful Conduct' (*Facebook* 2019) <https://www.facebook.com/communitystandards/hate_speech> accessed 26 June 2026.

305 Facebook (Chapter 3, n 304).

306 Facebook (Chapter 3, n 304).

307 Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* 4 (2018) 4.

308 Kaye (Chapter 3, n 285) 32–34.

309 Mark Zuckerberg, 'A Blueprint for Content Governance and Enforcement' (*Facebook* 15 November 2018) <https://www.facebook.com/notes/751449002072082/> accessed 11 June 2026.

310 Zuckerberg (Chapter 3, n 309).

311 Rebecca Bellan, 'Facebook Announces Members of Its Supreme Court' (*Forbes* 06 May 2020) <https://www.forbes.com/sites/rebeccabellan/2020/05/06/facebook-announces-members-of-its-supreme-court/#3759e11a3f12.> accessed 11 June 2026.

312 Bellan (Chapter 3, n 311).

313 Elizabeth Culliford, 'Facebook Names First Members of Oversight Board That Can Overrule Zuckerberg' Reuters (*Reuters* 06 May 2020) <https://www.reuters.com/article/technology/facebook-names-first-members-of-oversight-board-that-can-overrule-zuckerberg-idUSKBN22I2LQ/> accessed 11 June 2026.

only content that Facebook/Meta has taken down.[314] Further, the board lacks the authority to review Facebook/Meta's content moderation policies or procedures that guide removal decisions, including the algorithm that curates users' feeds and prioritises hateful content.[315] In addition, the board's rulings must comply with 'Facebook's content policies and values,'[316] which limits its ability to apply alternative sources of guidance, such as international human rights law.

336. The board's decisions are binding on just one case at a time and do not unambiguously establish a policy precedent for all similar cases, a power that would be essential to ensuring that the board's decisions would have a wider policy impact. Facebook/Meta must implement a board decision within a week,[317] unless the decision would be illegal according to national law. It is unclear how the board and Facebook/Meta navigates jurisdictional issues surrounding national laws on hate speech and disinformation.[318]

337. The board's authority is limited in both scope and speed. The board cannot review content on Facebook Dating or Marketplace, or messages sent through WhatsApp, Messenger, or Instagram Direct, all services owned by Meta.[319] Since its review process can take ninety days,[320] the board cannot act expeditiously to tackle viral content contributing to imminent public violence, as occurred during the atrocities against the Rohingya, which have been characterised as genocidal.[321] The bylaws make allowance for expedited review requests which might speed the process up, but 'expedited' can still mean up to thirty days.[322] Finally, the board can amend certain sections of the bylaws, but on the fundamental question of case review procedures (i.e. which cases are selected and how they are reviewed), Meta reserves its right to veto any board amendments it disagrees with.[323]

338. As a result of these substantial restrictions, some critics have dismissed the oversight board as 'toothless'[324] and inherently unable to address the source

314 Facebook, Inc, 'Oversight Board Bylaws' (2022) art 3.
315 Müller and Shwarz (Chapter 3, n 22) 27, 42.
316 Facebook, Inc – Bylaws (Chapter 3, n 314) art 1(3).
317 Facebook, Inc – Bylaws (Chapter 3, n 314) art 2 (2.3.1).
318 Jennifer Daskal, 'Speech across Borders' (2019) 105 Virginia Law Review 1605, 1607–1608.
319 Facebook, Inc – Bylaws (Chapter 3, n 314) 2 (1.2.1).
320 Facebook, Inc – Bylaws (Chapter 3, n 314) art 1(3.1) Case Review Procedures and Timeline.
321 HRC IIMM Report, Detailed Findings (Chapter 3, n 29) § 77, 1388–1441.
322 Facebook, Inc – Bylaws (Chapter 3, n 314) art 2(2.1.2).
323 Facebook, Inc – Bylaws (Chapter 3, n 314) art 5(1).
324 Josh Constine, 'Toothless: Facebook Proposes a Weak Oversight Board' (*TechCrunch* 28 January 2020) <https://techcrunch.com/2020/01/28/under-consideration/> accessed 12 June 2026.

of Facebook/Meta's content moderation problems, namely its 'attention economy'[325] business model that incentivises contentious and hateful speech to sell more advertising.[326] They have also portrayed the board as a half-measure to forestall government oversight and anti-trust investigations, an assertion that seemed vindicated in May 2020 when Meta launched an advocacy group called 'American Edge' to lobby in Washington against regulation of the tech sector.[327]

7.1 A Contextual Analysis of Online Speech

339. Given that social media users are dispersed around the globe and therefore not subject to the jurisdiction of any one country's laws,[328] and recognising the wide variation in national laws on freedom of expression, international human rights law can serve as the basis for social media companies' policies on content moderation.[329]

340. International law regarding speech regulation requires consideration of context. In his report to the UN General Assembly regarding online hate speech, then UN Special Rapporteur on freedom of expression and opinion David Kaye recommends that 'any enforcement of hate speech rules involves an evaluation of context and the harm that the content imposes on users and the public.'[330] In particular, context is essential to assessing the impact of speech.

---

325 Andrew Keen, 'The "Attention Economy" Created by Silicon Valley Is Bankrupting Us' (*TechCrunch* 30 July 2017) <https://techcrunch.com/2017/07/30/the-attention-economy-created-by-silicon-valley-is-bankrupting-us/> accessed 12 June 2026.

326 Dipayan Ghosh, 'Facebook's Oversight Board Is Not Enough' (*Harvard Business Review* 16 October 2019) <https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough> accessed 12 June 2026.

327 Tony Romm, 'Facebook Is Quietly Helping to Set up a New Pro-Tech Advocacy Group to Battle Washington' (*The Washington Post* 12 May 2020) <https://www.washingtonpost.com/technology/2020/05/12/facebook-lobbying-american-edge/> accessed 12 June 2026.

328 Statista, 'Leading Countries Based on Facebook Audience Sise as of April 2020' (*Statista* 2020) <https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/> accessed 30 April 2020; in 2020, only 189 million out of 1.623 billion (or 11%) of Facebook's daily active users were recorded as based in the United States and Canada: Facebook, Inc (Chapter 3, n 288).

329 Kaye (Chapter 3, n 281) §§ 5–14. *See* Evelyn Mary Aswad, '*The Future of Freedom of Expression Online*' (2018) 17 Duke Law and Tech Review 26, 42–57 applying international human rights law to the content moderation policies of social media companies.

330 Kaye (Chapter 3, n 281) § 58(d); *see also* Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham International Law Journal 939, 979 explaining that context is important in assessing the necessity of the limitation; Michael

Human rights law requires that any limits on speech be proportional to the legitimate ends that they serve,[331] and determining proportionality requires assessing speech in its context.[332] This principle is critical to the conversation about online hate speech: the need to evaluate expression in the context of its usage. In our understanding, the full meaning and potential effect of any speech act lies with the intent of the speaker, the content of the expression, and the context in which it is uttered.

341. Social media companies have global lists of words and expressions that are banned, and these lists are universal; that is, they generally do not vary according to country or region.[333] Highly local terms, such as the anti-Muslim racial epithet *kalar* that was repeatedly posted on Facebook by members of the Myanmar military during their campaign against the Rohingya, are often missed by content evaluation procedures either because they are in languages that are not particularly well-known in Western countries where social media companies are headquartered or because they are not on the list of universal hate speech terms.[334]

342. Alongside localised slurs, there is the problem of coded incitement and implicit threats that do not present as hate speech or violate hate speech guidelines, but are understood as incitements or threats by those with the requisite cultural knowledge. The campaign against the Rohingya in Myanmar, which has been characterised by the United Nations IIMM as genocidal, raised

---

Karanicolas, 'Squaring the Circle Between Freedom of Expression and Platform Law' (2020) 20 Pittsburgh Journal of Technology Law and Policy 177, 197 '[I]nternational courts have repeatedly found that assessing whether something is hate speech is an inherently contextual determination, based on whether the material was released under circumstances where it is likely to actually generate harm.'

331 Frank La Rue, 'Report of the Special Rapporteur on Freedom of Opinion and Expression' (2011) U.N. Doc. A/HRC/17/27 24.

332 The Rabat Plan of Action, a report produced in 2013 by a group of human rights experts convened by the UN High Commissioner for Human Rights, emphasised in particular the role of context in determining whether speech was so harmful as to require criminalisation; their factors required consideration of the 'social and political context,' the 'status of the speaker,' the '[e]xtent or reach of the speech act,' and the extent to which it was likely to result in harm. Kaye (Chapter 3, n 281) 13–14.

333 'The algorithms Facebook currently uses to remove hate speech only work in certain languages' in Billy Perrigo, 'Facebook Says It's Removing More Hate Speech than Ever Before. But There's a Catch' (*Time* 26 November 2019) <https://time.com/5739688/facebook-hate-speech-languages/> accessed 13 June 2026.

334 Perrigo (Chapter 3, n 333); see also *Application of the Convention on the Prevention and Punishment of the Crime of Genocide* (*The Gambia v Myanmar*) Application Instituting Proceedings and Request for the Indication of Provisional Measures (11 November 2019) ICJ 38–39, 44–45.

with particular urgency the importance of identifying coded or euphemistic speech threats. For example, the Independent International Fact-Finding Mission on Myanmar noted that it was difficult to identify online hate speech against the Rohingya because of '[s]ubtleties in the Myanmar language and the use of fables and allegories.'[335]

343. Coded speech is present in virtually all instances of genocide and mass atrocity. For instance, after the 1994 Rwandan genocide, the International Criminal Tribunal for Rwanda (ICTR) had to decide whether exhortations from leaders such as Jean-Paul Akayesu calling on Rwandans to go 'to work' and to fight and kill 'the *Inkotanyi*' (literally in Kinyarwanda, 'warriors') represented incitement to genocide against the Tutsi minority.[336] On the basis of expert testimony from Rwandan sociolinguist Mathias Ruzindana about the use of veiled incitement by Hutu Power extremists, the ICTR decided that both expressions did constitute direct and public incitement to exterminate Tutsis.[337] Social media posts by a modern-day Akayesu implicitly inciting harm by encouraging his followers to 'go to work' would likely not be removed under current content moderation policies because implicit speech requires sufficient contextual information to render its meaning evident.[338] Current content moderation policies do not integrate contextual factors into the decision-making and thus may permit coded calls to violence.[339]

8 Recommendation: a United Nations Digital Code of Conduct

344. Many observers have expressed unease with the fact that in the current arrangement, online speech is regulated by unaccountable private companies whose primary obligations are neither to voters nor to society at large, but to shareholders.[340] Rather than constrain this exercise of power, national governments are compounding it further by outsourcing their speech regulating

335 HRC IIMM Report, Detailed Findings (Chapter 3, n 29) § 1311.

336 Wilson (Chapter 2, n 34) 191.

337 Wilson (Chapter 2, n 34) 192.

338 Wilson (Chapter 2, n 34) 190–192.

339 Hamilton (Chapter 3, n 2).

340 See Luca Belli, Pedro Augusto Francisco and Nicolo Zingales, 'Law of the Land or Law of the Platform? Beware of the Privatisation of Regulation and Police' in Luca Belli and Nicolo Zingales (eds), *Platform Regulations: How Platforms Are Regulated and How They Regulate Us* (FGV Direito Rio 2017) 41, 53; Molly K Land, 'Against Privatized Censorship: Proposals for Responsible Delegation' (2020) 60 SSRN Electronic Journal 363, 399–415.

powers to private companies with little or no oversight.[341] For example, in 2017, Germany passed the pioneering Network Enforcement Law (Netzwerkdurchsetzungsgesetz – NetzDG)[342] that established measures imposing fines up to €50 million if platforms fail to remove speech that is 'manifestly unlawful'[343] under German law (including hate speech, defamation, and inciting speech) from their platform in as little as twenty-four hours.[344]

345. The German NetzDG has since been replaced with the EU Digital Services Act (DSA)[345] which entered into force in 2022 and which established a unified European framework for digital services and content moderation on online platforms. Except where it concerns content that is illegal under existing domestic or EU laws, the DSA does not specify the kinds of content which should be taken down from online platforms. It does, however, require increased transparency over the way in which content moderations are taken and how the algorithms employed by tech companies function. The rise of national and supranational regulatory frameworks represents a marked shift towards oversight and transparency over content moderation in online spaces, which has so-far been largely self-regulated by tech companies.

346. These developments highlight that one of the main alternatives to self-regulation by private companies is that states take on this regulatory function, but this brings pitfalls of its own. Scholars have long noted the paradox at the heart of human rights: states are the primary institutions upholding and enforcing human rights and states also the main violators of human rights. Therefore, we sound a necessary note of caution in discussions of online speech regulation. There is ample recent evidence to support the observation that censorious regimes frequently abuse incitement laws to stifle political and religious dissent. For instance, in 2019, the Chinese government convicted a Christian pastor for incitement of subversion of state power and sentenced

341 Land (Chapter 3, 340) 412–413.

342 The full name of the NetzDG is the Netzwerkdurchsetzungsgesetz. For a German-language version of the statute, see Beschlussempfehlung und Bericht des Ausschusses für Recht und Verbraucherschutz (6. Ausschuss) [Resolution and Report of the Committee on Legal Affairs and Consumer Protection (6th Committee)] (2017) Deutscher Bundestag: Drucksache [BT] 18/13013 (Ger.) <http://dip21.bundestag.de/dip21/btd/18/130/1813013.pdf> accessed 14 June 2026.

343 Deutscher Bundestag (Chapter 3, n 342) § 3(2).

344 Soraya Sarhaddi Nelson, 'With Huge Fines, German Law Pushes Social Networks to Delete Abusive Posts' (*NPR* 31 October 2017) <https://www.npr.org/sections/parallels/2017/10/31/561024666/with-huge-fines-german-law-pushes-social-networks-to-delete-abusive-posts> accessed 15 June 2026.

345 In Germany, the national Digital Services Act (Digitale Dienste Gesetz – DDG) published in 2023 implements the DSA and thereby effectively renders the NetzDG obsolete.

him to nine years in prison for saying that the Chinese Communist Party's ideology was 'morally incompatible with the Christian faith.'[346] In the same year, the Governor of South Dakota proposed new legislation banning 'incitement to riot' to stamp out protest by indigenous and environmental activists at the Keystone XL pipeline.[347] Around the world, defamation laws are used to target and silence environmental and human rights activists.[348]

347. At the same time, the international human rights system focuses on guiding and reforming state behaviour and relies, in large part, on state enforcement of agreed norms and rules. Thus, we must precisely articulate what measures are permitted and not permitted by states, and we encourage the international community to codify these agreed norms into a new UN Digital Code of Conduct, as indicated below.

348. First, we note that international human rights law is strongly supportive of freedom of expression, but permits limitations that are provided by law,[349] necessary to meet a legitimate objective,[350] and proportionate to the interest to be protected.[351] Article 19(3) of the International Covenant on Civil and Political Rights (ICCPR) states that the only objectives considered legitimate are:

a. Respect of the rights or reputations of others or
b. Protection of national security or of public order, or of public health or morals. International human rights law does not permit states to violate the rights of defenders by curtailing their legitimate democratic speech, by damaging their reputation, or by inciting violence against them.

Article 20(2) of the ICCPR explicitly prohibits 'any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.' Some forms of anti-human rights speech, including state-sponsored forms, are prohibited by international law, especially those that draw on

346 Eva Dou, 'Activist Chinese Pastor Gets Nine-Year Prison Sentence' (*The Wall Street Journal* 30 December 2019) <https://www.wsj.com/articles/activist-chinese-pastor-gets-nine-year-prison-sentence-11577698307> accessed 15 June 2026.

347 Stephen Groves, 'South Dakota Governor Plans Revision of Riot-Boosting Laws' (*AP News* 16 December 2019) <https://apnews.com/42f4533aac4913f2bb686d8d06c14b7b> accessed 15 June 2026; American Civil Liberties Union, 'South Dakota Governor Drops Anti-Protest Laws in Settlement Agreement with ACLU' (*ACLU* 23 October 2019) <https://www.aclu.org/press-releases/south-dakota-governor-drops-anti-protest-laws-settlement-agreement-aclu> accessed 15 June 2026.

348 OSCE Office for Democratic Institutions and Human Rights, 'Guidelines on the Protection of Human Rights Defenders' (2014) 70.

349 HRC, General Comment No. 34, Article 19 (Executive Summary, n 1).

350 HRC, General Comment No. 34, Article 19 (Executive Summary, n 1) §§ 28, 29.

351 HRC, General Comment No. 34, Article 19 (Executive Summary, n 1) § 34.

national, racial or religious tropes or that incite violence. Instead of making threats and inciting violence against human rights defenders, states have an obligation to take positive steps to protect freedom of expression, equal access to information, and freedom of association for their citizens.

349. States must cease and desist from targeting individuals and groups in their information operations and, under the auspices of the United Nations, negotiate, sign, and ratify a Digital Code of Conduct. The UN Digital Code of Conduct should, at a minimum, commit states to ensure transparency in their digital practices, require them to support freedom of expression and the exercise of basic citizenship rights online, and prohibit them from threatening and inciting violence against human rights defenders, and proscribe them from conducting unfettered surveillance and smear campaigns against their citizens.[352]

350. Correspondingly, States should consider whether to impose on technology companies a duty of care vis-a-vis their users, since they presently possess the principal capacity to regulate speech crimes. Social media companies currently detect and disrupt some state information operations, especially when they use networks of bots and engage in 'coordinated inauthentic behaviour,' but they do so in a globally variable and inconsistent manner.[353]

351. Platforms must move away from the current one-size-fits-all content moderation policy towards a context-specific approach that is informed by, and responsive to, the circumstances on the ground.[354] Content moderation policies must facilitate a pluralisation of speech norms and a decentralisation of their operations while conforming to international human rights principles of freedom of expression, transparency, accountability, and respect for due process.

352. Ideally, this means that human content moderation should be conducted by personnel who are native speakers of the language and who are familiar with the political and cultural context in which the speech is occurring. In at-risk countries where human rights are being violated on a mass scale or such a scenario is developing, social media companies are advised to open local offices and develop strong 'trusted partner' relationships with civil

352 See Kaye (Chapter 3, n 281) §§ 52–53 discussing the need for public oversight of surveillance technologies by states.

353 Cat Zakrzewski and others, 'How Facebook Neglected the Rest of the World, Fueling Hate Speech and Violence in India' (*Washington Post* 24 October 2021) <https://www.washingtonpost.com/technology/2021/10/24/india-facebook-misinformation-hate-speech/> accessed 15 June 2026.

354 Wilson and Land (Chapter 3, n 232) 1029; Wilson and Kiper (Chapter 3, n 232) 56.

society groups, including local journalists and human rights organisations with appropriate vetting to prevent groups with political affiliations instrumentalising this role to influence discourse on social media in their favour. Some companies have established internal working groups to monitor developments in countries with a heightened risk of political violence in order to assess how the local situation manifests itself on their platforms, but they need to go further in integrating the offline and online signals into their risk assessment matrix.

353. Some platforms conduct an expedited review of posts flagged by local trusted partners and this practice needs to be adopted more widely. In-country offices should be staffed with journalists, human rights attorneys, and political analysts who understand the political context, especially during elections when the risk of public violence is highest. Platforms should not rely significantly on Artificial Intelligence and automated content moderation alone but should integrate external signals identified by the local content and analysis teams into the content moderation matrix. Some companies include human rights activists as a protected category and wider adoption of this policy is warranted.[355] In EU Member States, trusted entities called 'trusted flaggers' – established based on a set of harmonised criteria – play a key role in implementing the DSA's strategy to address illegal online content.[356] These entities must function independently from the online platforms they are monitoring and follow established standards and procedures.[357] The trusted flagger mechanism does not prevent other individuals, private entities and civil society organisations from voluntarily collaborating with tech companies, but notices of illegal content made by DSA designated trusted flaggers must be treated with priority.[358]

---

355 Kari Paul, 'Facebook Rule Protects Journalists and Activists as "Involuntary" Public Figures' (*The Guardian* 13 October 2021) <https://www.theguardian.com/technology/2021/oct/13/facebook-involuntary-public-figures-journalists-harassment-bullying> accessed 10 April 2022; Facebook/Meta, 'Corporate Human Rights Policy' (*Facebook/Meta* 2021) § 4 Protecting Human Rights Defenders <https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf> accessed 10 April 2022.

356 EC, 'Trusted Flaggers under the Digital Services Act (DSA) | Shaping Europe's Digital Future' <https://digital-strategy.ec.europa.eu/en/policies/trusted-flaggers-under-dsa> accessed 15 June 2026.

357 EC (Chapter 3, n 356).

358 Council Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (DSA).

CHAPTER 4

Gathering and Submitting Evidence of Speech Crimes

354. Given the frequency of cases with speech acts dimensions in international criminal litigation, the collection and submission of admissible and verifiable evidence of such acts is a critical topic. The overarching framework governing evidence at the ICC is flexible and broadly worded. As speech acts are litigated more frequently, the standards governing the use of speech acts for criminal purposes will inevitably be developed in the jurisprudence and potentially in regulatory instruments. In the current era, evidence of speech crimes is often extracted from online sources, which is a dynamic evidentiary area in need of clarification. In the following sections, the guidelines explore the types of evidence of most relevant for this purpose, the means of identifying, archiving and preserving these materials, and evaluating and verifying such materials.

355. Chapter 6 of the Hartford Guidelines of 2017 examined the investigation of speech crimes, including the collection and analysis of evidence relevant to the locutionary, illocutionary, and perlocutionary aspects of a speech act. It was noted that speech crimes contrast to most other international crimes because they typically focus on public conduct, such as speeches and media statements, and that this made the remote collection of evidence particularly viable.[1] Those projections have proven prescient, as the prominence and use of social media has become ubiquitous. A growing array of actors have used evidence of speech acts collected remotely in order to demonstrate responsibility in legal proceedings, including UN investigative mechanisms,[2] international courts,[3] and university projects.[4] With the rapidly growth of online platforms, new risks have emerged, such as deep fakes and bot farms generating misleading information. The threat is real. For example, the Independent International Fact-Finding Mission on Myanmar found that derogatory posts were 'part of a systematic strategy used by the Myanmar military on Facebook

1 Wilson and Gillett (Key Guidelines, n 1) § 255.

2 HRC, Independent Investigative Mechanism for Myanmar, 'Anti-Rohingya hate speech on Facebook: Content and network analysis' (2024).

3 *Application of the Convention on the Prevention and Punishment of the Crime of Genocide in the Gaza Strip* (*South Africa v Israel*) Provisional Measures Order (26 January 2024) ICJ § 52.

4 Essex Human Rights Centre (Chapter 3, n 48).

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-006
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

to target the Rohingya population'.[5] Consequently, there is a need for this additional guidance regarding the collection and analysis of evidence of speech crimes in the age of social media.

1 Types of Evidence of Speech Crimes on Social Media

356. Speeches, broadcasts and inflammatory statements have frequently been entered into evidence at international criminal trials.[6] In the social media era, this type of evidence will only increase in significance, as a growing proportion of humans around the planet spend a substantial amount of their time on digital platforms. Notably, an ever-increasing number of people look to social media as a source of information and news and to participate in the "public square" of collective conversation, rather than traditional media sources. This creates the risk of harmful statements being propagated, especially as speech can be amplified online by humans or bots, for nefarious purposes.[7]

357. The specific social media platforms that are the most prominent and impactful when it comes to potentially inciting material are Facebook, Telegram, TikTok and Twitter/X, and to a lesser extent YouTube, Instagram, Reddit, and Spotify.[8] On these platforms, there is variation in the types of evidence that may demonstrate speech crimes. Items that could be used as evidence may be textual in nature, such as on Tweets/X posts; or composite pictures with accompanying statements, such as on Facebook or Instagram; or video, such as on TikTok or YouTube; or audio, such as on Spotify. The content may be expressed in virtually any language, as well as by the combination of words and meaningful images. Formats vary, with Twitter/X posts usually limited to 280 (up from their original 140) characters, TikTok videos ranging from 3 seconds to 60 minutes duration, while YouTube videos may be hours long. Many of the platforms enable live streaming, which usually involves a person

5 See Zarmsky (Chapter 3, n 1) 174 citing OHCHR, 'Report of the Independent International Fact-Finding Mission on Myanmar' (08 August 2019) A/HRC/42/50.

6 Lindsay Freeman and Raquel Vazquez Llorente, 'Finding the Signal in the Noise: International Criminal Evidence and Procedure in the Digital Age' (2021) 19 Journal of International Criminal Justice 163, 173.

7 Freeman and Vazquez Llorente (Chapter 4, n 6) 173.

8 See Zarmsky (Chapter 3, n 1) 178; Global Witness, 'Facebook, X/Twitter, YouTube and TikTok approve violent misogynistic hate speech adverts for publication in South Africa', (*Global Witness* 07 December 2023) <https://globalwitness.org/en/press-releases/facebook-xtwitter-youtube-and-tiktok-approve-violent-misogynisti-hate-speech-adverts-for-publication-in-south-africa/> accessed 26 June 2026; European Union Agency for Fundamental Rights, 'Online Content Moderation – Current challenges in detecting hate speech' (*FRA* 2023).

disseminating video and commentary during an event in real-time. There is variation across platforms in the availability to set up groups, and how accessible the content is within those groups to the wider public (with admission varying from fully controlled by a human administrator to effectively automated and available to anyone filling in a form). These distinctions can be important, particularly in relation to the public element of direct and public incitement to genocide.[9] Other technical features with forensic relevance include the fact that platforms vary in how they record the time a post is made and the extent to which they strip meta-data from an item when it us uploaded.

358. The views (and linked notion of 'impressions') demonstrate the reach of an item, which is a vital for assessing the impact of a social media post. For example, an Amnesty International investigation of speech crimes in the context of the Israel-Hamas conflict, which utilised analytical methodology set out in the Hartford Guidelines, included a variable called 'amplification' which essentially noted the number of views of the posts which were assessed as an additional factors relevant to their impact.[10] Views are a dynamic variable, which must be updated at key moments to avoid mistakenly under or over-estimating the reach of an item, particularly if it goes 'viral' after an investigation commences. Awareness of the unique features of different platforms, such as how they register and list views or impressions, as well as the types of digital evidence of speech crimes will be critical for its subsequent collection and use in international criminal proceedings.

359. On the question of amplification, it is instructive to refer back to the prior discussion on AI-driven algorithms which may amplify harmful content and increase engagement. For example, a user may post a statement, comment, meme, image, video or other content online with a limited reach and not intending it to go viral. If the content is then amplified through the platform's recommender algorithm, this can substantially increase its reach and potentially promote it to millions of other users. In this way, amplification through AI-driven algorithms may make the causal link between the original platform user's behaviour and the outcome of large-scale harm more ambiguous. However, it may be possible to establish foreseeability, a core concept in international criminal law in determining *mens rea*. The individual user who posts or shares inciting content nevertheless remains liable, as they can bear responsibility for outcomes they did not intend and a reasonable person could have foreseen that a post might have the potential to go viral.

---

9 On the 'public' element, see Wilson and Gillett (Key Guidelines, n 1) §§ 78–81 citing ICTR cases such as *Prosecutor v Nzabonimana* (Judgement) ICTR-98-44D-A (29 September 2014) §§ 231, 384.

10 See Amnesty International (Chapter 3, n 67) 33.

2 Obtaining, Archiving and Preserving Potential Evidence

360. Obtaining evidence of online speech crimes can be a challenge. In the ICJ case of *Gambia vs Myanmar*, Meta (Facebook) refused to provide evidence deemed private (such as Messenger messages between users) without a court order requiring it.[11] Although social media posts are typically public in nature, this does not mean they will continue to be publicly available in perpetuity. Like any commercial entity, social media companies and platforms can collapse or be discontinued (for example Skype, which was a major telecommunications company in the 2000s, has been discontinued since May 2025). They also typically remove content that violates community standards, which may constitute evidence of speech crimes.[12] Items can be removed manually by reviewers, whether on a proactive basis or in reaction to complaints raised by community members. Increasingly, items are also removed automatically via algorithms, and data retention policies vary widely. For most social media companies, it is unclear how this removed information is retained and how international tribunals may access it after a certain period. In this light, the need to preserve and archive digital evidence of speech crimes in a timely manner is particularly acute.

361. A more systematic archiving digital evidence of speech crimes can be achieved in multiple ways. The civil society organisation Witness assists frontline observers and witnesses to do so, by allowing them to upload captured footage relevant to any potential human rights violations, which could include inflammatory speeches and then storing the uploaded contents with protective geo and chrono tags to record when and where the item was created. The ICC itself is increasingly archiving large volumes of digital materials, and other entities such as the International, Impartial and Independent Mechanism (IIIM) for Syria have made significant advances in this respect. Less well-resourced entities, such as domestic non-governmental organisations (NGOs), are also storing digital materials. In so doing, they must pay particular heed to privacy and digital security requirements to ensure that they do not inadvertently leak sensitive materials which may reveal the identities of persons in vulnerable positions, including in countries subject to ongoing armed conflicts and societal repression.

362. When it comes to preservation of evidence that is potentially relevant to an international trial, social media companies have an important role to

11 See Zarmsky (Chapter 3, n 1) 178.

12 Maria de Arcos Tejerizo, 'Digital evidence and fair trial rights at the International Criminal Court' (2023) 36:3 Leiden Journal of International Law 753.

play in storing data. Where they remove materials from public view that may constitute or indicate speech crimes (or any other atrocity crimes), social media companies should preserve that material for investigative use. However, all too frequently materials are being removed from the internet, rendering them difficult to obtain by law enforcement.[13]

3 Evaluating Potential Evidence

363. The Hartford Guidelines noted the importance of analysing the cultural context of statements.[14] Understanding that cultural context when interpreting digital items is just as important as any other assessment of evidence of speech crimes. Evidence and assessments from historians, anthropologists, linguists, and sociologists can assist to decode expressions and appreciate their meaning. When assessing the cultural context of online speech acts, it is critical to recognise that content shared digitally is not limited to any one geographic location. A statement which might be understood by an audience in one location or cultural space in a certain way could have a completely different effect in another. In the online environment, there are many micro-cultures, with varying lexicons and means of communicating, which must be examined with an awareness of those cultural contexts in order to discern evidence of speech crimes and distinguish it from other communications and statements which may be offensive or shocking but fall within the penumbra of free expression.[15]

364. Alongside the cultural context, the language in which the speech is conveyed will need to be assessed and carefully translated into the working language of the assessor and potentially also for presentation to a Court or publicly for accessibility.[16] For evidence of speech crimes on social media platforms, the technological context is also important. Awareness of the technological context means assessing not only the contents of the statement, the identity of the speaker, the language or form of expression being used, and the identity of the audience, but also the format of the statement (text; images; video; etc …), the amplification of the statement (through shares, likes, replies, reposts, and other forms of drawing attention), the online presence of the speaker, and the social media consumption habits of the recipients of

13 Zarmsky (Chapter 3, n 1) 180.

14 Wilson and Gillett (Key Guidelines, n 1) § 255.

15 See, e.g. Meta Oversight Board, 'UK Drill Music' (*Meta Oversight Board* 22 November 2022) <https://www.oversightboard.com/decision/ig-pt5wrtlw/> accessed 25 June 2026.

16 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) § 14.

the information.[17] In general, to assess the speaker's online presence, significant factors include the social status of the alleged perpetrator, their network, political standing, position of authority, public reach to an audience, and history of posting on platforms.

365. Whereas the contents of online materials can be assessed in the same way as other speech acts, the digital era imports new technological contextual factors relevant to the evaluation of speech. For example, resharing an item could prima facie indicate support for the contents but could potentially also simply be an act designed to draw attention to it, possibly even with disapproval. Consequently, any additional information added when resharing will be critical.[18] Equally, the difference between adding a 'like' and an expression of shock (or other expression) will be salient in assessing whether a reshare is an endorsement or merely propagation. Certain memes have a history of supporting certain movements and so their etymological history should be assessed when determining their meaning. Links will also be relevant – if links are inserted to reputable news sources or academic journals this will indicate a neutral approach to the topic, whereas if links to extremist content, such as IS recruitment pages or Russian State-linked inflammatory figures such as Dimitri Medvedev, are inserted without caveats or warnings, this would be consistent with supporting speech crimes.

366. The evaluation of spoken words is facilitated by transcribing them. There are a growing number of platforms which transcribe materials automatically, such as the Ukraine War Archive for example, which allows users to upload materials to a database, where they are securely stored and can be searched while also providing automatic transcriptions of spoken text.[19]

367. In terms of substance, identifying potential items qualifying as speech crimes can be assisted by a framework. The Hartford Guidelines criteria have been used as a basis for parameters for the triage of potentially qualifying speech acts by multiple entities. This has typically involved taking the list of 14 factors set out in the Hartford Guidelines as indicative of speech crimes and adapting those to the specific circumstances.[20]

368. In its investigation of Russian influencers speech crimes in the context of Russia's invasion of Ukraine, FIDH used the following criteria, at least 3 of which had to be present to be tagged as potential persecution:

---

17 Suresha Perera, Nadeera Meedin, Maneesha Caldera, Indika Perera, Supunmali Ahangama, 'A comparative study of the characteristics of hate speech propagators and their behaviours over Twitter social media platform' (2023) 9:8 Heliyon 3.

18 Perera and others (Chapter 4, n 17) 3.

19 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) § 14.

20 Wilson and Gillett (Key Guidelines, n 1) § 320.

1. The intensity of the speech given its context, including the position of the speaker, means and frequency of dissemination and the overall environment; and at least two of the following content-related criteria:
2. Explicit or implicit calls for violence against members of the protected group;
3. Denigration of persons on the basis of their membership in the protected group;
4. Dehumanisation, verminisation, pathologisation, and demonisation of persons on the basis of their membership in the protected group;
5. Use of false and distorted narratives, such as conspiracy theories, false victimisations, and accusations in a mirror;
6. Choice of language, including the use of vivid images, graphic metaphors, euphemisms and derogative neologisms.[21]

369. Similar matrices have been used in investigations of speech crimes in the context of the Israel-Hamas conflict,[22] and other studies of Russian speech crimes during its invasion of Ukraine.[23]

4 Verification of Potential Evidence of Speech Crimes

370. In addition to evaluating the meaning and significance of speech acts in context, another analytical facet which is particularly important in the social media age is verifying that the item is authentic. As digital evidence has become more prevalent in international criminal proceedings,[24] so too has the need for verification. This essentially means checking that an item, and its contents, are what they purport to be. Verification is essential to verify that an item comes from who it purports to come from, and that it have not been manipulated in some manner so as to change its contents. A wholesale form of content manipulation is to stage it or artificially create it in the first instance.[25] One source of generated content is AI. For example, 'a deepfake video of Ukrainian

21 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) § 26.

22 See Amnesty International (Chapter 3, n 67) 33.

23 See Essex Human Rights Centre (Chapter 3, n 48).

24 Alexa Koenig states that the first Al-Werfalli warrant was a 'milestone, as it represented the first time the ICC cited videos pulled from social media – often referred to as open source evidence – as a basis for its charges'; Alexa Koenig, '"Half the Truth is Often a Great Lie": Deep Fakes, Open Source Information, and International Criminal Law' (2019) 113 American Journal of International Law Unbound 250, 250.

25 Başak Çalı and others, 'Evaluating digital open source imagery: A guide for judges and fact-finders' (*True Project* 2024), <www.trueproject.co.uk/osguide> 13.

President Zelensky calling his troops to surrender was circulated in 2022 on social media.'[26] Ensuring rigorous procedures are adopted to identify inauthentic content is essential for investigations of potential speech crimes.

371. Identifying the source of the item is called attribution. This is particularly important for speech acts, as mis-attributing a statement to the wrong person (whether due to intentional causes or by accident) could see them wrongly accused of acts they are not responsible for. Misattribution can also occur with items that depict real events, by relocating them to a different time or context.[27]

372. Verifying that the contents of an item of evidence have not been digitally altered is also critical. The contents of items can be cut, or augmented with filters and additional content such as captions, audio or images.[28] For example, opponents of the US Democratic Party created 'a real video of Nancy Pelosi, the Speaker of the House in the USA, that was edited to make it appear that she was intoxicated and slurring her words'. This was subsequently debunked, but not before it had spread widely through online social networks.[29] Altered metadata is another form of manipulating digital contents, which can be a means of misleading viewers as to its provenance. However, some changes to an item occur automatically and do not undermine its reliability as long as they are done transparently and consistently. For example, when uploading items to social media and sending them to interlocutors, the meta-data and/or format of the item is often changed. This does not automatically render the items inadmissible in any subsequent proceedings. In fact, the manipulation of items may be relevant to demonstrating responsibility for crimes, potentially as consciousness of guilt.[30] In all cases, investigators handling the items should be aware of these adjustments and factor them into their assessments.

373. Several challenges arise when seeking to verify digital open-source information. One is the sheer volume of materials; with several thousand X/Twitter posts generated every second and hundreds of hours of videos posted onto YouTube every minute.[31] This can make it difficult to locate specific items, and difficult to discern which is the original version of an item.

---

26 Bobby Allyn, 'Deepfake video of Zelensky could be 'tip of the iceberg' in info war, experts warn' (*NPR*, 16 March 2022) cited in Başak Çalı and others (Chapter 4, n 25) 13.

27 Başak Çalı and others (Chapter 4, n 25) 12.

28 Başak Çalı and others (Chapter 4, n 25) 12.

29 Başak Çalı and others (Chapter 4, n 25) 12.

30 On consciousness of guilt, see Matthew Gillett, 'Consciousness of Guilt' (*TedX Talks x University of Essex* 2023) <https://www.youtube.com/watch?v=KwozCUvoEnQ> accessed 27 June 2026.

31 Koenig (Chapter 4, n 24); Freeman and Vazquez Llorente (Chapter 4, n 6) 174.

374. Analysts have drawn our attention to the potential threat of deep fakes. Deep fakes are videos 'generated via algorithms that make it look like a person said or did something she did not'.[32] In seeking to verify such contents, it is important to maintain objectivity, check for editing or manipulation of the video, and then undertake further steps such as reverse image-search (to exclude the item actually being from an earlier conflict or period but being presented as relating to the present conflict).[33] Other steps include geo-location, chrono-location, checking the uploader's online history, assessing any narration, looking for visual indicia of authenticity, checking the meta-data, and seeking corroboration from other sources (triangulation).[34] In doing so, 'investigators [should] embrace 'multiple working hypotheses' about the credibility of visual and auditory content'.[35] There are also more technical solutions being developed. For example, 'InVID offers an online tool that allows verifiers to check the integrity of digital content to assess reliability.'[36] It is imperative to keep the safety and security of the reviewing team in mind at all times.[37]

375. In investigating potential speech crimes, it is imperative to identify and preserve digital evidence as quickly so as to ensure the best possible conditions for successfully prosecuting this conduct.[38] One tool to mitigate these risks is the use of Article 56 proceedings, whereby the Court can take measures in the face of a unique investigative opportunity, where time is of the essence to collect and preserve evidence which may otherwise become unobtainable.[39] At all times, quality control methods should be incorporated into investigations of speech crimes, such as those set out in the Berkeley Protocol on Digital Open Source Investigations.[40]

---

32 Koenig (Chapter 4, n 24).

33 Başak Çalı and others (Chapter 4, n 25) 21.

34 Witness, 'Video Evidence as a Field Guide' (*Witness* 2016) <https://vae.witness.org/video-as-evidence-field-guide/> accessed 28 June 2026.

35 (Chapter 4, n 24) (n 742).

36 (Chapter 4, n 24) (n 742).

37 See Wallace Fan, 'How to organise a collaborative OSINT project for litigation purposes: Takeaways from Project Tollgate' (*Essex Human Rights Centre* 2022) <https://hrcessex.wordpress.com/2022/05/26/how-to-organise-a-collaborative-osint-project-for-litigation-purposes-takeaways-from-project-tollgate/> accessed 26 June 2026.

38 Freeman and Vazquez Llorente (Chapter 4, n 6) 175–176.

39 Freeman and Vazquez Llorente (Chapter 4, n 6) 176.

40 Human Rights Centre UC Berkeley School of Law and UN OHCRC, 'Berkeley Protocol on Digital Open Source Investigations' (*Berkeley HRC* 2022) <https://www.ohchr.org/sites/default/files/2024-01/OHCHR_BerkeleyProtocol.pdf> accessed 27 June 2026.

5 Presenting Social Media Speech Crimes Evidence

376. Once evidence of potential speech crimes has been collected, it must be presented in judicial proceedings to establish facts and whether or not the accused is guilty. The following assessment focuses on the ICC, as this is the international court with the broadest jurisdictional reach, potentially encompassing the entire globe (in the case of a UN Security Council referral under Article 13(b) of the Rome Statute).

377. As a general approach, the ICC adheres to the approach of other international courts in its broadly framed evidentiary frameworks.[41] This is reflected in the key Rome Statute provisions, such as Article 69(3) (titled 'Evidence'), which that '[t]he Court shall have the authority to request the submission of all evidence that it considers necessary for the determination of the truth'; Article 69(4), which allows the Court to take into account, *inter alia*, the probative value of the evidence and 'any prejudice that such evidence may cause to a fair trial or to a fair evaluation of the testimony of a witness';[42] as well as Rule 63(2), which states that the Judges may 'freely' assess the evidence.

378. There are restrictions in the Court's evidentiary framework, the most relevant of which are: Article 69(7), which provides that evidence will be inadmissible if obtained by means of a violation of the Statute or internationally recognised human rights, where that would cast 'substantial doubt' on the reliability of the evidence or would be 'antithetical' to, and seriously damaging of, the integrity of the proceedings; and Article 69(8), which prevents the Court

41 See *Prosecutor v Thomas Lubanga* (Corrigendum to Decision on the admissibility of four documents) ICC-01/04-01/06-1399-Corr (21 January 2011) § 26 concluding that the Court 'enjoys a significant degree of discretion in considering all types of evidence.'; Ken MacLean, 'Interactive Digital Platforms, Human Rights Fact Production, and the International Criminal Court' (2023) 15:1 Journal of Human Rights Practice 84; Reinhold Gallmetzer, 'The Trial Chamber's Discretionary Power To Devise The Proceedings Before It And Its Exercise In The Trial Of Thomas Lubanga Dyilo', in Carsten Stahn and Göran Sluiter (eds), *The Emerging Practice of the International Criminal Court* (Brill 2009) 501–524, 507; Riccardo Vecellio Segate, 'Cognitive Bias, Privacy Rights, and Digital Evidence in International Criminal Proceedings: Demystifying the Double-Edged AI Revolution' (2021) International Criminal Law Review 243.

42 The fact that evidence is deemed admissible does not determine the weight, which will ultimately be afforded in the court's judgement; see *Prosecutor v Jean-Pierre Bemba Gombo* (Decision on the admission into evidence of materials contained in the prosecution's list of evidence) ICC-01/05-01/08-1022 (19 November 2010). Some chambers have simply allowed parties to submit evidence without contemporaneously ruling on its admissibility, saving the consideration of the evidence until the conclusion of proceedings.

from ruling on the application of a State's national law when deciding on the relevance or admissibility of evidence collected by a State. Finally, a provision that is particularly relevant where the verification and interpretation of speech is involved is Rule 64(2), which requires that '[a] Chamber shall give reasons for any rulings it makes on evidentiary matters.'

379. Given the increasing reliance on digital forms of speech acts by international courts,[43] the Court's guidance on digital open-source information is pertinent. In ICC jurisprudence, chambers recommend that the tendering party should indicate the date and location from which open-source materials were obtained.[44] The best evidence rule is also relevant,[45] as this requires that parties collect and submit the best version of the available evidence in the circumstances.[46] For items such as posts or videos for example, this means that the original version of the item should be tendered rather than re-posts, unless no longer possible or otherwise inappropriate in the circumstances (such as if the focus was on the re-sharing by an accused rather than an original post by a different person). More technically, if the statement is captured in digital form, chambers should seek to verify the item's provenance, meta-data, authorship,[47] authenticity,[48] and address any potential tampering.[49]

380. The concept of hearsay is a particularly significant issue for speech crimes cases. At the ICC, as with other international courts, hearsay is not prohibited and so is presumptively admissible, subject to the usual rules of procedure and evidence. However, in recent years, chambers have indicated caution in relation to hearsay materials.[50] In particular, open-source materials based

43 See *South Africa v Israel*, Provisional Measures Order (Chapter 4, n 3) § 52.

44 *Prosecutor v Germain Katanga* (Decision on the Prosecutor's Bar Table Motions) ICC-01/04-01/07 (17 December 2010) § 24(a) and (d).

45 See Richard Glover, *Murphy on Evidence* (15th edn, Oxford University Press 2017) Section 2.5.1.

46 *Prosecutor v Vidoje Blagojević and Dragan Jokić* (Decision on the Admission into Evidence of Intercept Related Materials) IT-02-60-T (18 December 2003) § 25; *Prosecutor v Milan Martić* (Decision Adopting Guidelines on the Standards Governing the Admission of Evidence) IT-95-11-T (19 January 2006) § 7.

47 See Nikita Mehandru and Alexa Koenig, 'ICTs, Social Media, & the Future of Human Rights' (2019) 17 Duke Law and Technology Review 129, 135.

48 *Bemba et al.* TC (Chapter 1, n 50) § 222.

49 See Stefano Trevisan, 'Open-source information in criminal proceedings: lessons from the International Criminal Court and the Berkeley Protocol' (2021) Giurisprudenza Penale Web 10-4. International Criminal Court case law on evidence derived from online open-source information.

50 See, e.g. *Prosecutor v Ali Muhammad Ali Abd-Al-Rahman ('Ali Kushayb')* (Public redacted version of 'Defence response to Prosecution's second request to submit material from the bar table of 31 January 2023') ICC-02/05-01/20-871-Conf (13 February 2023) 'Although

on anonymous hearsay have been frequently denied admission, including news articles,[51] NGO reports, and even United Nations reports.[52] The risks with hearsay evidence include the fact that the original statement is made without the maker being under an oath to tell the truth, and outside there courtroom thereby depriving the opposing party of the ability to cross-examine on issues of recollection, perception, sincerity and meaning, as well as an inability to observe the statement maker's demeanour.[53]

381. The most direct means to present evidence of speech crimes is through a bar table motion. That means requesting the Court, through written motion, to receive the evidence as part of the evidentiary collection on which the judges of a trial chamber can base its decision. Bar table motions have been used regularly in international criminal trials, including at the ICTY, ICTR, and ICC.[54] Submitting evidence of speech crimes from online sources could be achieved through bar table motions, as long as sufficient explanation is provided regarding the provenance, reliability and relevance of the items. However, given the relative novelty of submitting online evidence of speech crimes, it is likely that tendering parties would seek to do so through witnesses wherever possible. The witnesses may be persons who created or disseminated the online items, such as bloggers and posters, or summary type witnesses, such as investigators who can explain where the items were found, how they were processed, and what they relate to.

---

there is no prohibition on hearsay before the Court, there is an inherent risk in this type of evidence. The Chamber may therefore take such risks into consideration when attributing the appropriate probative value to items of evidence consisting mainly or exclusively of hearsay'. *The Prosecutor v. Laurent Gbagbo and Charles Blé Goudé* (Decision adjourning the hearing on the confirmation of charges pursuant to article 61(7)(c)(i) of the Rome Statute) ICC-02/11-01/11-432 (03 June 2013) §§ 28–35.

51 *Abd-Al-Rahman* (Chapter 4, n 50) Defence response § 37 'It is anonymous hearsay which the Defence cannot possibly test in any meaningful way. The strength of the purported linking evidence to the Accused's nickname is so weak and of such low probative value that its formal recognition would be outweighed by its prejudicial effect'.

52 Lindsay Freeman, 'Prosecuting Atrocity Crimes with Open Source Evidence', in *Digital Witness: Using Open Source Information for Human Rights Investigation Documentation and Accountability* (Sam Dubberley, Alexa Koenig, & Daragh Murray eds., 2019), 50 citing *Prosecutor v Jean-Pierre Bemba Gombo* (Decision on the admission into evidence of items deferred in the Chamber's 'Decision on the Prosecution's Application for Admission of Materials into Evidence Pursuant to Article 64(9) of the Rome Statute') ICC-01/05-01/08-2721 (27 June 2013).

53 See Demetra Sorvatzioti, 'Free Evaluation of Evidence: Does the ICC need a Law of Evidence?' (2021) 22 International Criminal Law Review 895, 905.

54 *Katanga* (Chapter 4, n 44) § 24(a) and (d).

382. Nonetheless, given the complex technical nature of collecting and interpreting evidence of speech crimes, it is likely they would also need to be accompanied by expert evidence from specialists with advanced knowledge and skills relating to online and digital technology. The ICC's rules governing the privileged role of experts are relatively minimalist, in keeping with its open approach to the admission of evidence. The Court's jurisprudence has specified that, unlike 'lay' witnesses,[55] expert witnesses are permitted to provide their opinions, which are not restricted to events they directly experienced through their senses.[56] This can be important for interpreting speech acts, as experts are allowed to explain why inferences can be drawn based on events.[57] This also differs from 'summary' witnesses, such as police officers, investigators, and other officials who may be called to summarise collections of reports for example, but who are not proposed as persons of particularly specialised skill in the manner of an expert.[58]

383. For courts other than the ICC, many of these same considerations will apply. However, there are significant differences, such as the openness to hearsay evidence, which diverges between common law systems where hearsay is presumptively disallowed subject to (often numerous) exceptions and civil law systems where it is presumptively permissible along with any relevant materials gathered in the dossier for the case. Investigators and attorneys looking into potential speech crimes cases should closely assess the applicable legal frameworks in the jurisdictions where they might litigate these matters.[59]

---

55 *Bemba et al.* TC (Chapter 1, n 50) § 20. See also Déirdre Dwyer, *The Judicial Assessment of Expert* Evidence (Cambridge University Press 2008) 78.

56 Karen MacGregor Richmond and Sebastiano Piccolo, 'Between Fact and Opinion: The Sui Generis Approach to Expert Witness Testimony in International Criminal Trials' (2021) 22 International Criminal Law Review 1016, 1017.

57 Artur Appazov, *Expert Evidence and International Criminal Justice* (Springer 2016) 19.

58 See, e.g. *Bemba et al.* (Transcript of Hearing) ICC-01/05-01/13-T-11-Red-ENG (30 September 2015) oral decision in which the Chamber notes Prosecution's submission that the witness will 'summarise evidence' and holds that 'the Chamber understands the anticipated testimony of this witness to be about facts that are immediately apparent in the evidence collected by the Prosecution and are available without any specific knowledge beyond what is contained in this material. (…) As long as the testimony is limited in this way, the Chamber does not see a difficulty with the witness's testimony' 6.

59 See, generally, Matthew Gillett and Wallace Fan, 'Expert Evidence and Digital Open Source Information: Bringing Online Evidence to the Courtroom' (2023) 21:4 Journal of International Criminal Justice 661–693.

6 Fair Trial Concerns

384. In gathering and submitting evidence of speech crimes, investigators must pay close attention to fair trial protections, particularly the equality of arms principle. That is captured in Article 67 of the Rome Statute, which provides that an accused person has the right 'to a fair hearing conducted impartially, and to the following minimum guarantees, *in full equality*' (emphasis added) and is consistent with the requirement in Article 21(3) that the application of the Court's legal framework must be consistent with internationally recognised human rights, which includes the right to equality of arms as captured in Article 10 of the Universal Declaration of Human Rights ('Everyone is entitled in full equality to a fair and public hearing by an independent and impartial tribunal, in the determination of his rights and obligations and of any criminal charge against him.') and Article 14 of the International Covenant on Civil and Political Rights ('In the determination of any criminal charge against him, everyone shall be entitled to the following minimum guarantees, in full equality').

385. In speech crimes cases, the rights of particular relevance are those under Article 67(1)(b) ('To have adequate time and facilities for the preparation of the defence and to communicate freely with counsel of the accused's choosing in confidence'); (e) ('To examine, or have examined, the witnesses against him or her and to obtain the attendance and examination of witnesses on his or her behalf under the same conditions as witnesses against him or her. The accused shall also be entitled to raise defences and to present other evidence admissible under this Statute') and 67(2) on the disclosure of exculpatory evidence ('defence evidence in the Prosecutor's possession or control which he or she believes shows or tends to show the innocence of the accused, or to mitigate the guilt of the accused, or which may affect the credibility of prosecution evidence').

386. In practice, these rights are substantial, as the Office of the Prosecutor will typically have the competence and established institutional standing to approach social media companies to obtain evidence, as well as access to training on the identification, extraction, and interpretation of speech crimes evidence.

387. The principle of equality of arms also comes into play where AI systems are used as a tool to process and analyse evidence. AI tools developed by private actors may be used in evidence analysis without clear insights into how they are designed or how they are trained to make decisions or

assessments.[60] A report by Fair Trials on the regulation of artificial intelligence for use in EU criminal justice systems stressed that evidence generated by such systems may be impossible to verify or replicate, particularly where defendants are denied access to the source code, training data, or decision-making pathways of the tools being employed.[61] This lack of transparency compromises and undermines the principle of equality of arms, preventing defendants from meaningfully contesting the evidence presented against them by the prosecution.[62] This risks resulting in a profound asymmetry where the prosecution may rely on technologically enhanced tools to bolster its case, while the defence is not provided with transparent information about how such tools function and is therefore unable to effectively challenge their validity.

60 Taylor R Moore, 'Trade Secrets and Algorithms as Barriers to Social Justice' (*Center for Democracy & Technology* 2017).

61 Bruno Min and Griff Ferris, 'Policy Paper: Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU' (*Fair Trials* 2022) 24.

62 Min and Ferris (Chapter 4, n 61) 3.

CHAPTER 5

Universal Jurisdiction Cases and Speech Crimes

388. Universal jurisdiction is becoming an increasingly important avenue for the prosecution of international crimes especially in Europe, but also in the United States. In universal jurisdiction cases, states assume jurisdiction over crimes committed outside of their territory and without the victim or the perpetrator being a national of the prosecuting state. In certain types of universal crimes with jus cogens status in international law (such as slavery, genocide or torture), it is accepted that the offence may be prosecuted by any competent tribunal, and indeed, said tribunals have an *erga omnes* obligation to prosecute.[1] As such, universal jurisdiction navigates a tension between state sovereignty and the need for effective prosecution of serious crimes and is therefore typically reserved for serious offences of international nature.[2] Universal jurisdiction may remedy gaps left by the limited reach of the ICC. Importantly, it enables the prosecution of 'lower level' perpetrators.

389. The use of universal jurisdiction as an avenue to combat impunity for international crime is increasing. The 2024 Universal Jurisdiction Annual Review published by TRIAL International[3] counted 36 new cases which were either opened or made public in 2024 and 27 convictions in first instance or appeal.[4] These figures appear indicative of a steadily upward trend since 2022, which 'was marked by an unprecedented mobilisation of existing and judicial resources to respond to the international crimes committed in Ukraine'.[5] For comparison, for 2023 TRIAL International counted 36 new investigations

1 Cherif M. Bassiouni, 'International Crimes: 'Jus Cogens' and 'Obligatio Erga Omnes'' (1996) 59:4 Law and Contemporary Problems 63–74.

2 On universal jurisdiction generally see Anne Lagerwall and Marie-Laurence Hébert-Dolbec, 'Universal Jurisdiction' in *Max Planck Encyclopaedias of International Law*, last updated July 2022.

3 In collaboration with Civitas Maxima, the Center for Justice and Accountability (CJA), the European Center for Constitutional and Human Rights (ECCHR), the International Federation for Human Rights (FIDH) and REDRESS.

4 Trial International, Civitas Maxima, CJA, ECCHR, FIDH and REDRESS, 'Universal Jurisdiction Annual Review: New Developments in 2024' (*Trial International* 08 April 2025) <https://trialinternational.org/latest-post/universal-jurisdiction-annual-review-new-developments-in-2024/> accessed 26 June 2026.

5 Trial International, Civitas Maxima, CJA, ECCHR, FIDH and REDRESS, 'Universal Jurisdiction Annual Review 2023' (*Redress* November 2023) 10 <https://redress.org/storage/2023/04/01_TRIAL_UJAR_2023_DIGITAL_27_03.pdf> accessed 27 June 2026.

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-007
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

opened, and 16 convictions and for 2022 TRIAL International counted 12 countries prosecuting universal jurisdiction cases and 23 convictions.[6] The number of cases under investigation is more difficult to pinpoint, given that the relevant crimes are often investigated as part of large scale investigations covering numerous individuals and incidents. According to Eurojust, in 2022 more than 6000 investigations of cases of core international crimes were closed or still pending in EU member states alone.[7] Against the backdrop of current armed conflicts such as Russia's war against Ukraine, the number of cases and the importance of universal jurisdiction is expected to increase further.[8]

390. In principle, speech can be prosecuted domestically via universal jurisdiction as crime or as a mode of liability, similar to the practices of international courts and tribunals. However, the attempt of so doing may face several challenges. Although domestic legislation on international crimes may be modelled after international law, the successful prosecution of speech acts may ultimately hinge on national investigative tools, available resources (both human and financial), laws, and doctrines.

391. The following sections illustrate the existing law and jurisprudence on speech elements in universal jurisdiction cases. They focus on differences compared to international courts and tribunals, highlighting the challenges of prosecuting speech as a crime or as mode of liability via universal jurisdiction. The chosen examples stem from jurisdictions active in the practice of universal jurisdiction: Belgium, Germany, and the Netherlands. The general challenges and opportunities in investigating and proving speech crimes abroad that apply for international courts and tribunals apply to universal jurisdiction, too.

1 Domestic Prosecution of Speech Charged as a Crime

392. In parallel to the law and practice of international courts and tribunals, domestic Courts may consider speech acts as crimes. While there have been successful prosecutions of propaganda activities under domestic laws, e.g. prohibiting the support of terrorism, cases pertaining to speech acts as core international crimes are rare.

6 Trial International and others, UJAR 2023 (Chapter 5, n 5) 13.

7 Eurojust, 'At A Glance: Universal Jurisdiction in EU Member States' (*Eurojust* 2023) <https://www.eurojust.europa.eu/sites/default/files/assets/at-a-glance-universal-jurisdiction-in-eu-member-states.pdf> accessed 28 June 2026.

8 Eurojust (Chapter 5, n 7).

2 Domestic Prosecution of Direct and Public Incitement to Genocide

393. In the absence of proceedings for direct and public incitement to genocide in universal jurisdiction cases, the following discussion reviews the existing legislation that may be relevant to incitement to genocide cases. Stated plainly, domestic legislation may not criminalise direct and public incitement to genocide in the same way as the Rome Statute and yet provide promising avenues of accountability for such acts.

394. Some states allow for the prosecution of direct and public incitement largely mirroring the Rome Statute. For example, in the Netherlands, public incitement to genocide is covered by Article 3(2) of the International Crimes Act (*Wet Internationale Misdrijven*) of 2003 and carries the same penalty as attempted genocide. Here, universal jurisdiction applies, if the suspect is identified and present in the Netherlands.[9]

395. Within the EU, and thus in some of the most active universal jurisdiction states, criminalisation of severe forms of incitement are further mandated by the Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law.[10] The implementation of the Council framework decision by the member states varies.

396. The experience of Germany illustrates the (limited) potential of addressing direct and public incitement via a law based on the Council Framework Decision. Under German law, incitement to the core international crimes is covered by a provision in the 'ordinary' criminal code. Since a 2022 amendment the law explicitly prohibits publicly condoning, denying or grossly trivialising the crimes in a way that is likely to incite hatred or violence against the protected groups or persons belonging to these groups and thus disturbing public peace.[11] This provision thus goes further than the Rome Statute as it is not limited to incitement to genocide but covers the other core crimes, too. However, its additional requirement of a disturbance of public peace is limiting.

397. In addition, the Code of Crimes against International Law (Völkerstrafgesetzbuch – VStGB) incorporates the core international crimes of the Rome

9 Wet Internationale Misdrijven (International Crimes Act) 2003 (Netherlands) art 3(2).

10 Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law [2008] OJ L328/55.

11 StGb n (Chapter 2, n 228) s130(5).

Statute into domestic German law. Sections 1 and 6 through 12 of the Code of Crimes against International Law confer true universal jurisdiction on German courts, enabling the investigation, prosecution, and trial of suspects irrespective of their nationality, that of the victims, or the location of the alleged crimes.[12]

3 Domestic Prosecution of War Crimes and Crimes against Humanity through Speech Acts

398. Speech acts as an element of war crimes have featured in at least one universal jurisdiction case.[13] Degrading speech was relevant in a case concerning alleged war crimes committed by a radical Islamist group in Syria in 2015, prosecuted in the Netherlands.[14] The Hague District Court convicted Ahmad al-Y for co-perpetration of violation of Common Article 3 of the Geneva Conventions, consisting of outrages upon personal dignity, in particular humiliating and degrading treatment of persons who have been placed *hors de combat*.[15] The conviction was based on an incident in which the accused, according to the findings of the Court, amongst other degrading acts 'stood and/or posed next to [...] deceased persons and called the aforementioned (deceased) persons 'dogs' and 'carcasses of al-Assad' and sang songs.'[16] The Court of Appeals however acquitted the accused of this charge, concluding that despite its being disrespectful, the acts did not meet the threshold for outrages upon personal dignity, taking into consideration that the persons concerned were deceased.[17]

399. As speech can be disseminated online, prosecutions of core international crimes away from the place of atrocities do not always need to take recourse to universal jurisdiction. In some cases, the accused may have disseminated speech supporting or endorsing atrocities from another state. For

12 Völkerstrafgesetzbuch (VStGB – Code of Crimes against International Law) 2002 (Germany) s1, s6–12.

13 The Hague District Court, Case Number: 09/748011-19 (21 April 2021) § 5.4.3.; The Hague Court of Appeals, Case Number: 2200128321.v (21 April 2022) § 9.3; see also *Public Prosecutor's Office v Ahmad al-Y* (Appeal Judgement) Case Number: ECLI:NL:GHDHA:2022:2858; 2200128321.v (06 December 2022).

14 *Ahmad al-Y* Appeal (Chapter 5, n 13).

15 The Hague District Court, Case Number: 09/748011-19 (Chapter 5, n 13) § 5.4.3.

16 The Hague District Court, Case Number: 09/748011-19 (Chapter 5, n 13) § 5.4.3.

17 The Hague Court of Appeals, Case Number: 2200128321.v (21 April 2022) (Chapter 5, n 13) § 9.3; *Ahmad al-Y* Appeal (Chapter 5, n 13).

example, The Hague District Court also convicted Yousra L. for the war crime of outrage upon personal dignity for disseminating IS video material with degrading comments on Telegram.[18] However, this case was not a matter of universal jurisdiction, as the accused committed the relevant acts when present in the Netherlands.

400. Speech acts may also constitute a basis for domestic charges of crimes against humanity, particularly in the form of persecution. For example, at the Colombian Special Jurisdiction for Peace (*Jurisdicción Especial para la Paz* (JEP)), speech acts have been part of the charges for persecution in the case against ex-Congressman and politician Luis Fernando Almario Rojas.[19] Notably, the JEP cited to the Hartford Guidelines to support this charge.

4 Domestic Prosecution of Speech as a Mode of Liability

401. Speech elements may further feature as mode of liability in universal jurisdiction cases. Here, the domestic legal standards for the respective mode of liability are key. When speech is involved as mode of liability, domestic courts are likely to refer to the modes of liability established in domestic law, rather than in international criminal law. The vast majority of jurisprudence in domestic courts, which shapes the standards required for the distinct modes of liability, typically developed in response less large-scale conflicts. In the area of speech crimes, where questions such as the required nexus between speech and harm are key, the application of domestic notions of modes of liability might lead to diverging results.

18 The Hague District Court, Official Dutch Summary (29 June 2021) 09/748012-19 <https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2021:6620&showbutton=true&keyword=09%252f748012-19-P&idx=2> accessed 27 June 2026; Eurojust, 'National Jurisprudence summary: Dutch case no. 09/748012-19 and 09/748012-19-P' <https://www.eurojust.europa.eu/sites/default/files/inline-documents/29.06.2021._netherlands_district_court_summary.pdf> accessed 27 June 2026.

19 Although this is not an exercise of universal jurisdiction per se, it does provide an example of a domestically established body prosecuting hate speech as a part of the charge of the crime against humanity of persecution. See JusticeInfo: 'Colombia, Special Tribunal Indicts First Prominent Politician', 22 June 2023, https://www.justiceinfo.net/en/118357-colombia-special-tribunal-indicts-first-prominent-politician.html accessed 30 June 2026.

4.1 Belgium: Perpetration of War Crimes and Genocide

402. Speech featured, although not as the central aspect, in a 2019 case in Belgium. A Rwandan national was convicted to 25 years imprisonment by a Belgian Court for perpetrating war crimes and genocide in Rwanda in 1994.[20] The Court found that the accused reported Tutsis to soldiers, who then killed them.[21] In so doing, according to the Court, the accused provided help to the soldiers.[22] Without his assistance, these crimes could not have been committed.[23] Therefore, the Court found him guilty of having committed war crimes in the form of killings.[24] Under Article 66 of the Belgian Criminal Code, someone who provides help without which the crime could not have been committed is punished for perpetrating (and not only aiding and abetting) such crime.[25]

403. Additionally, the Belgian Court highlighted the role of the accused as running a militia and supporting the messages of authorities at a public meeting, thereby enabling the massacres against Tutsis.[26] The Court referred to statements of the accused reported by witnesses to establish his genocidal intent.[27]

4.2 Germany: Aiding and Abetting War Crimes

404. A case against a Rwandan citizen in Germany concerning war crimes committed by the *Forces Combattantes Abacunguzi* (FOCA) militia in the Democratic Republic of Congo represents an unsuccessful attempt to

---

20 Court of Assizes of the Brussels Capital Administrative District (Decision in NF) FD30-98.102/02 (6th Session of 2019); Eurojust, Unofficial translation of Court of Assizes of the Brussels Capital Administrative District FD30-98.102/02 6th Session of 2019' <https://www.eurojust.europa.eu/sites/default/files/Documents/pdf/19.12.20_belgium_cour_d_assises_criminal_Judgement_anonymised_complete.pdf> accessed 27 June 2026.

21 Court of Assizes of the Brussels Capital Administrative District (Chapter 5, n 20) Unofficial Translation 9.

22 Court of Assizes of the Brussels Capital Administrative District (n 800) Unofficial Translation 10.

23 Court of Assizes of the Brussels Capital Administrative District (Chapter 5, n 20) Unofficial Translation 10.

24 Court of Assizes of the Brussels Capital Administrative District (Chapter 5, n 20) Unofficial Translation 10.

25 Code Pénal Belgique (Belgian Penal Code) 1867, as amended 2014 (Belgium) art 66.

26 Court of Assizes of the Brussels Capital Administrative District (Chapter 5, n 20) Unofficial Translation 15.

27 Court of Assizes of the Brussels Capital Administrative District (Chapter 5, n 20) Unofficial Translation 20.

prosecute speech as aiding and abetting.[28] The case concerned attacks against civilians carried out by the FOCA Militia belonging to the FDLR in the east of the Democratic Republic of Congo in 2008 and 2009. The accused Dr. Ignace Murwanashyaka acted as political leader and highest representative of the FDLR.[29] As such, he carried out propaganda and publicity activities for the FDLR.[30] For example, he supported militia fighters in their activities through messages expressing gratitude[31] and oversaw propaganda on a website, motivating the fighters.[32]

405. In 2015 Higher Regional Court of Stuttgart found that these propaganda activities contributed to war crimes being committed by the militia, as the fighters knew that the political leadership of the organisation would deny the acts and that therefore, they did not have to fear consequences.[33] The accused used press statements to deny wrongdoing of the organisation, despite knowing that alleged acts occurred, and blamed political opponents for the ongoing violence.[34] The Court was satisfied that his doing amounted to 'psychological aiding and abetting' (*'psychologische Beihilfe'*):[35] the propaganda activates increased the likelihood of the militia continuing to commit war crimes and lowered the likelihood of the activities ceasing.[36]

406. The Federal Court of Justice overturned the conviction regarding the psychological aiding and abetting through propaganda.[37] The Court found that the Lower Court had not established a sufficient nexus between the

28 See ECCHR, 'Groundbreaking trial in Germany Rwandan FDLR rebel leaders sentenced' (*ECCHR* 2015) <https://www.ecchr.eu/en/case/groundbreaking-trial-in-germany/> accessed 27 June 2026; see also Friedrich-Alexander Universität Rechts- und Wirtschaftswissentliche Fakultät and International Nuremberg Principles Academy, 'Case Information Sheet: Prosecutor v Dr. Ignace Murwanashyaka'(2018) <https://vstgb-datenbank.de/media/pdfs/CIS_Mustervorlage_DrIgnaceMurwanashyaka_AR_deutsch_KB.pdf> accessed 27 June 2026.

29 *Kriegsverbrecherprozess wegen Straftaten im Bürgerkrieg in der Demokratischen Republik Kongo: Einstufung der Forces Démocratiques de Libération du Rwanda als eine terroristische Vereinigung im Ausland – Dr. Ignace Murwanashyaka* (Judgement) Higher Regional Court of Stuttgart (28 September 2015, file no. 5-3 StE 6/10, § 4.

30 *Kriegsverbrecherprozess Murwanashyaka* (Chapter 5, n 29).

31 *Kriegsverbrecherprozess Murwanashyaka* (Chapter 5, n 29) §§ 306, 335.

32 *Kriegsverbrecherprozess Murwanashyaka* (Chapter 5, n 29) § 399.

33 *Kriegsverbrecherprozess Murwanashyaka* (Chapter 5, n 29) § 337.

34 *Kriegsverbrecherprozess Murwanashyaka* (Chapter 5, n 29) § 342 et seq.

35 *Kriegsverbrecherprozess Murwanashyaka* (Chapter 5, n 29) § 1829 et seq.

36 *Kriegsverbrecherprozess Murwanashyaka* (Chapter 5, n 29) § 1830.

37 Federal Court of Justice Germany (Judgement) file no. 3 StR 236/17 (20 December 2018) § 102.

propaganda and the decision of the militia to commit the concrete acts in question.[38] Neither could a general upholding of the morale and creating the impression of impunity meet the threshold of psychological aiding and abetting.[39]

407. In its reasoning on this point, the Court referred exclusively to the jurisprudence of domestic courts rather than international courts and tribunals. Throughout universal jurisdiction cases, German courts commonly draw from the jurisprudence of international tribunals regarding the interpretation of elements of international crimes. However, for modes of liability not specific to international criminal law, such as aiding and abetting, the understanding modes of liability are based on domestic law and practice. Therefore, the extent to which speech elements – in particular in the form of propaganda activities – can be successfully prosecuted, may depend on the jurisdiction.

5 Cross-Cutting Issues in Universal Jurisdiction Cases

408. Regardless of whether speech is charged as a crime or features as one or more mode of liability, the investigation and prosecution of the crimes, as well as the representation of victims come with challenges distinct to speech crimes. There is a need for translation services as well as the involvement of experts in often times rare languages. One of the key risks arises from domestic authorities lacking knowledge of the language and context in a given conflict situation. This issue is particularly acute where speech crimes are concerned. It is therefore crucial that procedures and procedural law allow for the involvement of the required language experts, including socio-linguistic experts.

409. A key feature of universal jurisdiction is that it promotes increased accountability for serious international crimes and gross human rights violations by enabling their prosecution in domestic courts. Specifically, regarding speech crimes, however, there is a risk that national authorities may overlook, misinterpret or underestimate speech due to a lack of knowledge of cultural context and meaning of given phrases. While dedicated international courts or tribunals overseeing a specific contextual situation may have specific resources (both human and financial) dedicated to interpretation and translation throughout all phases of proceedings, including at crucial points throughout the investigations, this may be more challenging in a universal jurisdiction

38 Federal Court of Justice Germany (Chapter 5, n 37).
39 Federal Court of Justice Germany (Chapter 5, n 37).

case – especially if the languages in question are those of lesser diffusion or those based on oral traditions.

410. In this light, the role of NGOs in monitoring, intervening in (as amici curiae) and potentially even submitting cases to national authorities is potentially significant. NGOs that intervene in multiple jurisdictions on international crimes cases can provide information to the authorities to ensure that speech elements are not overlooked during investigations and prosecutions, as well as ensuring that human rights freedoms, such as that of liberty of expression, are duly upheld.

411. For investigations, open-source information can play a key role and assist when access to relevant territory or potential witnesses is restricted. For example, there have been cases prosecuted based on social media photos focused on charges of the war crime of outrages against personal dignity.[40] Because speech often occurs across social media platforms in current times, the use of open source investigations has considerable potential in this context, both in making the evidence discoverable and in overcoming challenges of being unable to access certain countries or areas that may be relevant to the investigation.

412. As always, the rights of the accused should not be compromised. Some jurisdictions may not be compliant with international standards, thus undermining the legitimacy of the legal process. In absentia proceedings, for example, give rise to concerns. Due process and fair trial rights should be examined and upheld.

413. Ultimately, for universal jurisdiction cases, domestic authorities should closely observe 'floor (and ceiling)' requirements that the universal jurisdiction prosecutions of speech crimes should adhere to. The minimum is internationally recognised human rights standards, such as freedom of expression and due process rights. The higher level consists of best practices regarding preservation of evidence, cooperation and the integration of the latest technology in analysis and presentation of evidence of speech crimes.

40 Eurojust, *'Prosecuting War Crimes of Outrage upon Personal Dignity Based on Evidence from Open Sources: Legal Framework and Recent Developments in the Member States of the European Union'* (*Eurojust* 2018) <https://www.eurojust.europa.eu/sites/default/files/assets/2018-02-prosecuting-war-crimes-based-on-evidence-from-open-sources-en.pdf> accessed 11 June 2026.

CHAPTER 6

Direct and Public Incitement to Aggression, Crimes against Humanity, and War Crimes

414. As yet, the potential preventative function of the legal regulation of speech remains unrealised. There was only one article relating to prevention in the 1948 UN Convention on The Prevention and Punishment of Genocide and that established direct and public incitement to genocide. The Convention and subsequent case law comprehended incitement to genocide as an inchoate crime in which the conduct is the crime and no further consequences need occur for it to be suppressed by law. Confusion has entered into the jurisprudence as a result of Article 25 of the Rome Statute of the International Criminal Court, which positions direct and public incitement to commit genocide in the Statute alongside modes of liability for completed crimes.[1] This creates significant ambiguity as to whether the prosecution of direct and public incitement to genocide under the Rome Statute must be linked to the physical commission of genocide. In seemingly converting incitement from an inchoate stand-alone crime to a contingent mode of liability and potentially requiring the commission of genocide before a prosecution can be mounted, such a stance consequently undermines the preventative value of incitement as an inchoate crime. To strengthen this potential preventative function of international criminal law, Article 25(3)(e) of the Rome Statute should be expanded to include inciting any of the Rome Statute crimes, as formulated in the suggested amendment below. This should explicitly confirm the inchoate nature of incitement, not requiring an atrocity crime to occur in order to prosecute its intentional, direct and public incitement as follows:

> Intentionally, directly, and publicly incites others to commit any of the crimes contained in the Statute, thereby significantly increasing the likelihood of their occurrence. For the purpose of this provision, it is not necessary that the incited crime(s) be committed or attempted.[2]

415. The proposed amendment enhances and emphasises the potential preventative function of international criminal law through several means.

1 Wilson and Gillett (Key Guidelines, n 1) § 4.
2 Wilson and Gillett (Key Guidelines, n 1) § 101.

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-008
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

Importantly, it seeks to address the current ambiguity within the Rome Statute around the status of direct and public as an inchoate crime,[3] or as a mode of liability contingent upon a genocide occurring by confirming the inchoate status of incitement to genocide.[4] It clarifies that it should not be necessary to wait for genocide to physically occur to prosecute those who are inciting towards and advocating for its commission. Furthermore, it extends the application of Article 25(3) to *any* of the international crimes for which the ICC has jurisdiction in accordance with the Rome Statute, thereby enhancing the potential preventative function of international criminal law towards the commission of crimes against humanity, war crimes and (potentially) the crime of aggression.

416. The Hartford Guidelines set out the justifications for this formulation, placing particular emphasis on its preventive function:

> The rallying cry, 'never again', is heard after each successive genocide. With mankind's capacity to cause death and destruction greater than ever before, the importance of preventing genocide, and other serious group-based violence, before it manifests is self-evident. Nonetheless, the preventive function of international criminal law remains underdeveloped.
>
> Direct and public incitement to genocide is a case in point. When the Contracting Parties to the Genocide Convention of 1948 included direct and public incitement to genocide in its terms, they sought to target incipient efforts to foment group-based hostility that could evolve into violent genocidal acts. Whereas the Genocide Convention clearly provides that direct and public incitement is an inchoate head of liability and does not require the completion of genocide, the Rome Statute of the ICC does not appear to replicate this approach. In the Rome Statute, direct and public incitement is located in article 25, with the modes of liability rather than the crimes. Because of this placement, it is unclear

3 As it has been treated by ad hoc tribunals. See also Wilson and Gillett (Key Guidelines, n 1) § 103.

4 For more on direct and public incitement against genocide, see Susan Benesch, 'Vile Crime or Inalienable Right: Defining Incitement to Genocide' (2008) 48:3 Virginia Journal of International Law 486–528. The ICTR has clearly stated that direct and public incitement to genocide is an inchoate crime: see *Nahimana et al.* AC (Chapter 3, n 167) § 678. However, under the Rome Statute (Key Guidelines, n 2), direct and public incitement to genocide has been included in the provision on modes of liability (as art 25(3)(e)), introducing some ambiguity regarding the nature of direct and public incitement to genocide. In Nahimana, the ICTR Appeals Chamber repeated the established legal position that direct and public incitement to genocide is an inchoate crime, but the outcome of the case nonetheless placed considerable weight on the fact that the inciting speech acts and the genocide itself were contemporaneous.

whether genocide need actually be committed in order for direct and public incitement to be charged. If it is necessary to wait for genocide to occur before its proponents and advocates can be prosecuted, then the preventive function of this form of liability will be significantly curtailed.

The following proposal therefore seeks to address the most serious forms of unlawful speech. We propose an amendment to article 25(3)(e) of the Rome Statute, which reads as follows:

> *Intentionally, directly, and publicly incites others to commit any of the crimes contained in the Statute, thereby significantly increasing the likelihood of their occurrence. For the purpose of this provision, it is not necessary that the incited crime(s) be committed or attempted.*

At present, article 25(3)(e) of the Rome Statute refers only to direct and public incitement of genocide. The proposed amendment would see a form of liability[5] included in the Rome Statute covering not only those persons who urge others to commit genocide, but also those who call for crimes against humanity, war crimes and (potentially) the crime of aggression. It would remove the current anomaly whereby direct and public calls for crimes such as extermination, rape or torture are not criminalised per se.

The proposed amendment would redress the current ambiguity in the formulation of direct and public incitement to genocide in the Rome Statute,[6] which has created confusion as to whether it is an inchoate crime (as the ad hoc tribunals considered) or a mode of liability requiring genocide to actually occur, as indicated by its placement in article 25 and the lack of wording to the contrary.[7]

Importantly, the proposed amendment confirms the inchoate nature of this form of liability. This would strengthen the Court's preventive function, as the

5 Direct and public incitement is here referred to as a 'form of liability' rather than a 'mode of liability' as it differs from the traditional conception of modes of liability as contributions to crimes that occur. In this sense, the proposed direct and public incitement formulation is similar to the form of liability of attempt under art 25(3)(f) of the Rome Statute (Key Guidelines, n 2).

6 The current formulation and placement of direct and public incitement to genocide in the Rome Statute could lead to confusion. The ICTY (Chapter 2, n 195) and the ICTR (Chapter 2, n 195) both included direct and public incitement within their jurisdictions (art 4(3) and art 2(3), respectively) and treated it as an inchoate crime. Whereas it appeared settled that direct and public incitement to genocide is an inchoate crime under customary international law, the Rome Statute (Key Guidelines, n 2) includes direct and public incitement in art 25, with the modes of liability rather than the crimes. This placement contributes to ambiguity about whether genocide need actually be committed in order for direct and public incitement to be charged. If it is necessary to wait for genocide to occur before its proponents and advocates can be prosecuted, then the preventive function of this form of liability will be significantly curtailed.

7 Wilson and Gillett (Key Guidelines, n 1) §§ 98–103.

direct and public incitement could be prosecuted without waiting for the execution of the atrocity. This contrasts with its more restricted jurisdiction over soliciting and inducing crimes under article 25(3)(b), for example, which require that the crime either occurs or is attempted (i.e. the perpetrator began to execute the crime but was thwarted by circumstances). Where direct and public calls are made for the commission of atrocities, the international community should not have to wait, like an ambulance at the bottom of a cliff, for the violence to manifest before measures can be taken against those urging such crimes.

The proposed amendment makes it necessary to show that the incitement 'significantly' increased the likelihood of genocide. This filter is designed to exclude less serious speech acts, such as fanciful calls for crimes or statements by persons with no real possibility of prompting anyone to commit grave crimes. The term 'significantly' is well-known to international lawyers, and can thus benefit from the guidance of international case law. The term 'significance' is also used in the social sciences in statistical tests of correlation and causation the social science literature.[8] In this manner, the proposed amendment can be interpreted in light of existing legal and academic sources of guidance. To avoid any doubt, the proposed amendment explicitly states that the incitement must be intentional.

Several legal sources provide support for the direct and public incitement of atrocity crimes, including:

- The ICCPR, article 20(1) ('Any propaganda for war shall be prohibited by law') and article 20(2) ('Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law').
- CERD, article 4: 'States Parties 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, with due regard to the principles embodied in the Universal Declaration of Human Rights and the rights

8 For modes of liability, the international courts have used various formulations to describe the threshold of the contribution that the accused needs to make for the crime to be held liable, including the terms 'substantial' and 'direct'. These terms may also help inform the analysis of whether the incitement in question has significantly increased the likelihood of the crime occurring. For example, the term 'substantial' has been considered to be a higher threshold than 'significant' at the ad hoc tribunals. Consequently, if conduct were considered to have substantially increased the likelihood of the crime occurring, then it would *a fortiori* meet the 'significant' threshold.

expressly set forth in article 5 of this Convention, inter alia: (a) Shall 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 assistance to racist activities, including the financing thereof; (b) Shall declare illegal and prohibit organizations, and also organized and all other propaganda activities, which promote and incite racial discrimination, and shall recognize participation in such organizations or activities as an offence punishable by law; (c) Shall not permit public authorities or public institutions, national or local, to promote or incite racial discrimination'.

– The Genocide Convention, article 3(c) ('The following acts shall be punishable … direct and public incitement to commit genocide') and article 5 ('The Contracting Parties undertake to enact, in accordance with their respective Constitutions, the necessary legislation to give effect to the provisions of the present Convention, and, in particular, to provide effective penalties for persons guilty of genocide or any of the other acts enumerated in article III').

At a time when extreme violence is so prevalent and the fires of discriminatory hatred so easily stoked; it could not be more essential to enhance the preventive function of international criminal law. Enacting the proposal set out above would be a measured but firm step towards realising the law's potential.

417. There is a strong normative, doctrinal, historical, and moral basis for the extension of Article 25(3)(e) to cover incitement of any of the atrocity crimes set out in the Rome Statute.[9]

418. In light of the full-scale Russian invasion of Ukraine in February 2022, which was preceded by numerous inflammatory statements by significant Russian media figures (and others), FIDH and other organisations concluded that 'mass atrocities against Ukrainian civilians, have been made possible by years of intense and escalating propaganda of hate designed to justify and facilitate Russia's territorial conquest of parts or all of Ukraine, and the subjugation or removal of civilians who oppose becoming part of the "Russian World".[10] Similarly, the Commissioner-General of UNRWA and Amnesty International

9 This is also outlined and explained in the Wilson and Gillett (Key Guidelines, n 1) §§ 97–107.

10 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41) Executive Summary, 2 see also §§ 88, 280 for further context.

found that the long-standing pattern of dehumanising and derogatory discourse by Israeli officials towards the Palestinian population had legitimised and normalised violence, compounding the crisis in Gaza.[11] These examples further demonstrate the need for liability for direct and public incitement to any Rome Statute crimes, as proposed in the Hartford Guidelines in 2018.

11 Amnesty International (Chapter 3, n 67) 236; and UNRWA (Chapter 3, 86).

CHAPTER 7

Key Guidelines for the Assessment of Speech Crimes under International Criminal Law

Based on the foregoing discussion, we propose the following Key Guidelines:

1. Speech acts that may incur criminal liability should be evaluated in the context of their usage. Analysing speech acts can be facilitated by conducting a two-step process: first, the linguistic interpretation of the meaning of the statement(s) made within the relevant context; and, second, the legal assessment of the significance of the statement(s), and particularly any contribution that the statement(s) made to the alleged crimes.[1]
2. Governments and digital companies removing unlawful content should ensure that evidence of speech crimes is preserved. Actors researching and investigating speech crimes linked to social media should verify content incorporating up-to-date technological tools and recording their investigative steps.
3. Analysing alleged speech crimes linked to social media requires taking account of the number of views and amplification of the speech act and the speaker's socio-political role, position, authority, and reach.
4. Disinformation is multi-faceted as it can directly contribute to crimes and can also distract from offences and thereby contribute to a climate of impunity or mislead the international community. Temporally, disinformation prior to (incitement), during (facilitating), or after (cover-up) crimes should be investigated.
5. Assessments of alleged disinformation as a contribution to atrocity crimes should address whether (i) the suspect had knowledge of the crimes (whether planned or committed); (ii) the suspect intentionally and knowingly disseminated untrue information; and (iii) the suspect did so to further the criminality.
6. Speech crimes charged under international criminal law must meet a gravity threshold to qualify for prosecution. If it is charged as a mode of liability, the speech act must satisfy the requisite contribution threshold

1 These consider the 14 factors identified in the Hartford Guidelines relating to the status of the speaker, the content of the message and the context of the speech act) form part of the second step of this process. See Richard Ashby Wilson and Matthew Gillett, *The Hartford Guidelines on Speech Crimes in International Criminal Law* (Peace and Justice Initiative 2018) § 320.

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-009
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

under that mode to the crime (or crimes) which meets the gravity threshold. The speech act should be more than a remote occurrence in the sequence of events leading up to the crime. If the speech itself is alleged to constitute the crime, it must be of sufficient seriousness to meet the requirements of that crime. In applying this threshold, courts should take into account the nature of the speech act, its repetition, its reach, and its impact.[2]

7. In universal jurisdiction cases in domestic systems involving speech acts, these recommendations should be incorporated into national laws, doctrines, and investigative methods. All cases prosecuted under universal jurisdiction should adhere to internationally recognised due process and human rights principles.
8. Corporate responsibility should be extended to speech crimes and should be implemented in line with these recommendations. Criminal responsibility should only be applied where the relevant gravity threshold is met. Other regulatory sanctions such as retractions, fines, restriction of licence, and dissolution can be considered for alleged speech crimes which do not meet the gravity threshold (and could also be imposed along with criminal responsibility where appropriate), though always bearing in mind the fundamental right of free expression. Natural persons using corporate structures to commit speech crimes should be held accountable irrespective of the availability of corporate responsibility mechanisms.
9. To strengthen the preventative function of international criminal law, States should affirm the obligation to prevent speech crimes. At the international level, this goal should be pursued by adopting an amendment to Article 25(3)(e) of the Rome Statute to include inciting any of the Rome Statute crimes, as formulated below. The new article confirms the inchoate nature of incitement by not requiring an atrocity crime to occur in order to prosecute its intentional, direct and public incitement:

> Intentionally, directly, and publicly incites others to commit any of the crimes contained in the Statute, thereby significantly increasing the likelihood of their occurrence. For the purpose of this provision, it is not necessary that the incited crime(s) be committed or attempted.

2 This Key Guideline does not seek to add an element to any crime but rather highlights the relevance of the gravity admissibility requirement under the Rome Statute. See *Rome Statute of the International Criminal Court* (adopted 17 July 1998, entered into force 01 July 2002) 2187 UNTS 3, art 17(1)(d).

10. States, with the assistance of civil society, and other relevant entities should implement these recommendations. State Parties to the Rome Statute should consider the amendment to Article 25(3)(e) and propose it to the Assembly of State Parties for referral to the Working Group on Amendments. The drafters of these Guidelines stand ready to assist with the consideration and implementation of the proposed amendment.

1 Updating the Risk Assessment Model for Speech Crimes

419. The risk assessment model for speech crimes developed in the Hartford Guidelines has been used as the basis for the investigation and analysis of speech crimes by several civil society organisations. This includes the 'Communication submitted under Article 15 of the Rome Statute of the International Criminal Court – Situation in Ukraine: Hate Speech as the Crime Against Humanity of Persecution' drafted by the FIDH and the report 'You Feel Like You Are Subhuman: Israel's Genocide Against Palestinians in Gaza' by Amnesty International.[3]

420. The below risk assessment model has been updated to include aspects specifically pertaining to online speech and the use of social media or other digital platforms to promulgate speech crimes. The following conditions relating to the status of the speaker, the content of the message and the context of the speech act predict the causal effects of inciting speech. Because speakers may use veiled language, the analysis requires linguistic expertise and should consider the social, historical, political and military context.[4] These factors raise the probability and imminence that a speech crime occurs. However, not all elements need to be present to heighten the risk of international crimes. As such, the factors in the risk assessment model should not be treated as a checklist but rather as an aid in the case-by-case evaluation of speech, which must take into account all the factors holistically within the overarching context.

1.1 Risk Assessment Model for Speech Crimes in the Digital Age

1. The speaker holds a recognised position of authority or social standing and has significant influence through media channels. This can include governmental figures, authority figures within a political organisation,

3 FIDH, Kharkiv Human Rights Protection Group, and Center for Civil Liberties (Chapter 3, n 41); Amnesty International (Chapter 3, n 67).

4 For a risk assessment model for incitement, see Wilson and Kiper (Chapter 3, n 232).

non-State actors, religious leaders, or individuals with other forms of social prestige such as online influencers. It may also include such individuals who, though they may not necessarily hold a high position of authority, have established access to propaganda networks;

2. The speaker, or their channel or platform, is perceived as credible by their audience. The term platform should be considered in a broad manner and may include traditional media like newspapers, television or radio broadcasts, podcasts, and also digital platforms such as social media, online messaging apps, chatrooms and online fora, gaming servers etc.;
3. The speaker is perceived as charismatic by their audience and the channel or platform used to disseminate their message has widespread distribution;
4. The speaker is particularly skilful in their use of pre-existing cultural symbols and narratives to cultivate historical grievances;
5. The speaker uses dehumanising utterances, refers to past atrocities and calls for revenge against the out-group;
6. The speaker resorts to intense language, replete with vivid images, graphic metaphors and exaggerations;
7. The speech, or their channel or platform, is perceived as 'powerful' by the audience;
8. The speaker's message is widely disseminated through mass communication fora, particularly social media platforms;
9. The speaker wields a monopoly on the means of communication or has the capacity to censor and suppress information or drown out counter-narratives through their reach and amplification. The restrictions imposed on the freedom of expression exceed those permissible under Article 19(3) ICCPR for one or more of the four reasons: i) they are not provided by domestic law; ii) they do not have a legitimate aim; iii) they are not necessary; iv) they are not proportionate;
10. The emotional state of the audience is influenced and/or more prone to manipulation by circumstances of insecurity and uncertainty, or by the format of the social media platform;
11. There is a history of intergroup conflict between the in-group and out-group;
12. There is a national or local political election in the next 12 months, or there was a contested national or local election in the previous 12 months;
13. There is significant polarisation of political parties along religious, ethnic, or racial lines;
14. The number of instances of inter-group violence has increased in previous months, and the instances have gone unpunished;

15. The speech instils fear by labelling a perceived direct or indirect threat and then suggesting a distinct and foreseeably violent course of action that can be taken by the audience to remove the source of that threat.

1.1.1 Factors to Consider Pertaining Specifically to Social Media:

- Where the speaker's message is widely disseminated through the use of social media or other digital platforms:
  - How many times has the speaker's message or content been liked, shared, and commented on?
  - Is there any indication as to the natural or manufactured origin and reach of the message/content, for example, does it appear that bot farms have been used to manipulate and artificially inflate engagement metrics or drive amplification?
  - Has the same message been repeated using similar wording and key terms across multiple posts?
  - Is there any unnatural pattern of posting or in the comments on a post?
  - Has the message or content been cross-posted to multiple platforms?
  - Are there any symbols, emojis, GIFs or other digital images and icons which appear to have been used as coded language to spread propaganda, hide explicit hate speech, call for violence etc., and evade content moderation algorithms?
  - Is there any evidence that the speaker's message has been picked up beyond social media?
- There are controlled and intentional shutdowns of telecommunication systems as a result of which the speaker wields a monopoly on the means of communication or has the capacity to censor and suppress information. Counter-narrative content is deleted or access to independent media, which may serve as a counter to the spread of disinformation, is restricted. The speaker controls the narrative or can influence the platform's content moderation policy to quell dissent.

CHAPTER 8

Concluding Remarks

421. These PJI Guidelines confront the new digital reality in which societies operate. Digital platforms enable users to convey meaning and provoke responses and actions in listeners instantly and globally. Yet speech does not escape scrutiny and potential legal consequences just because it is conducted online. To the contrary, legal frameworks must be applied with equal force in online environments. The virtual context can have profound impacts on the way in which speech is disseminated, amplified, and interpreted. New AI technology has exacerbated this trend, and is increasingly taking human words and amplifying, contorting, and misattributing them across a range of virtual environments. Moreover, AI is capable of generating harmful speech in and of itself, with specific human prompting.

422. These new circumstances present challenges to the traditional legal paradigm of speech crime analysis and to the investigation and adjudication of such harms. Approaches to investigations, evidence, and adjudication need to be reconceptualised to address the emerging technologies, and the normative and procedural framework governing international criminal law (whether applied before international bodies or in domestic courts through universal jurisdiction) requires re-examination. Against that backdrop, the PJI Guidelines explore key topics surrounding speech crimes in the digital era, with a view to expanding on the Hartford Guidelines and adding an innovative body of research to the literature on speech crimes, which is also designed to be of practical use for litigants, regulators and civil society.

423. In Chapter 1, the analysis underscores the fact that the interpretation of speech crimes – particularly those conveyed through indirect, implicit, or veiled language – turns significantly on how the intended audience understands the message within its broader social, political, and cultural environment. The analysis of the persuasive or mobilising force of such speech will commence with its literal wording but will then place significant emphasis on the meaning attributed to it by its recipients in context. Where the intended harmful effect is embedded in the communication, even in coded or implicit form, the speaker may incur legal responsibility, provided the requisite elements of the crime and any applicable mode of liability are established. In the jurisprudence on direct and public incitement to genocide, the Trial Chamber in *Akayesu* emphasised that the decisive inquiry is whether the audience would immediately comprehend the import of the message. This reinforces

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-010
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

the principle that interpretation is inseparable from context. At the same time, contextual analysis does not supplant careful scrutiny of the specific content and legal effect of each order. The existence of an armed conflict, for example, did not shield Mladić's statements from potential criminal liability. These decisions further confirm that individuals who issue orders do so from positions of authority; their elevated status amplifies both the impact of their words and the scope of their potential responsibility.

424. These doctrinal insights acquire particular significance in the digital era. Speech acts now frequently occur on social media platforms, where communication may take the form of written posts, emojis, videos, audio recordings, reposts, 'likes,' or other forms of engagement. Although English is increasingly prevalent, in the digital era it is expressed in hybridised forms that incorporate these emojis, acronyms, slang, memes, and multimodal combinations of text, audio, and visual symbolism. In such environments, it becomes especially important to maintain a clear distinction between linguistic interpretation (a factual inquiry into meaning) and judicial evaluation (a normative assessment of criminal responsibility). Conflating these dimensions risks allowing legal conclusions to shape factual interpretation, or vice versa, thereby undermining analytical clarity and legal certainty.

425. Moreover, digital ecosystems foster sub-cultures that generate specialised vocabularies and coded expressions that are often unintelligible to outsiders. In cases centred on speech acts within communities, expert evidence should extend beyond traditional linguistic analysis to include familiarity with the relevant digital subculture and its communicative practices. As in earlier jurisprudence, the circumstances of each case will remain decisive in guiding interpretation. Yet the pervasive and contemporaneous nature of online communication introduces an additional evidentiary layer. Public posts issued alongside operational directives, for example, whether for explanatory or propagandistic purposes, may illuminate how orders are framed, justified, or understood in real time.

426. Meaning is also contingent upon temporal and spatial factors. However, in the digital era, 'time' and 'place' can no longer be confined to a physical setting. They may extend to virtual domains – including immersive online spaces and the so-called metaverse – alongside traditional geographic locations. The relevant interpretive framework may therefore encompass digital platforms populated not only by human users but also by virtual representations and increasingly autonomous AI systems.

427. Chapter 2 sets out the pernicious and enduring role of disinformation in the provocation of atrocity crimes. Surveying cases from the post-World War II proceedings against Fritzsche and Streicher, to the ICTY cases of *Mladić*,

Karadžić, and Gvero (in *Popovic et al.*), to the ICC cases of *Mbarushimana* and *Ruto and Sang*, the analysis traces a line throughout the history of international criminal law up to present times showing the reoccurring role of speech in initiating crimes. The chapter also shows the relevance of societal circumstances in the settings of Rwanda and Germany, where human rights interpretations are influenced by the historic experiences of genocide in those settings. Disinformation and propaganda would most likely be prosecuted as the mode of liability of aiding and abetting. However, the provision containing aiding and abetting at the ICC (Article 25(3)(c) of the Rome Statute) is phrased in a potentially highly restrictive manner. If interpreted and applied strictly, the purpose test in Article 25(3)(c) is likely to hinder the prosecution of professional propagandists, such as Mbarushimana, under a theory of aiding and abetting. In the digital era, the ramifications will be far-reaching. Persons who significantly and knowingly contribute to crimes through the spread of disinformation will nonetheless potentially escape liability by pointing to differing purposes other than criminal, such as monetary, political, or religious for example, so as to preclude the application of aiding and abetting liability.

428. Chapter 3 addresses how disinformation and hate speech have long been transmitted through traditional media channels such as radio, television, and print journalism. In the contemporary information environment, however, social media platforms have assumed a central role in shaping how such speech is produced, amplified, and received. These platforms enable the rapid, transnational dissemination of harmful narratives on an unprecedented scale, significantly altering both the reach and the impact of inciting and threatening speech. As a result, social media has become a powerful vector through which messages that encourage or normalise violence can circulate widely and with little temporal delay, reaching audiences that may include direct or indirect perpetrators of physical harm.

429. At the same time, the digitalisation of communication has transformed the evidentiary landscape of international criminal accountability. Content generated and shared on social media can document crimes as they unfold, preserve traces of intent and coordination, and provide access to situations of violence that might otherwise remain obscured. This dual character of social media – as both a facilitator of harm and a potential repository of evidence – creates new opportunities for accountability while also introducing complex legal, technical, and institutional challenges.

430. Chapter 3 situates these developments within a broader historical and doctrinal context. It begins by revisiting the pre-digital case of *Nahimana et al.*, in which an international tribunal found that incendiary broadcast media played a decisive role in inciting genocide. Against this backdrop, the chapter

then turns to a series of contemporary case studies that illustrate the increasingly prominent role of social media in the context of mass atrocity crimes. These include the mistreatment of the Rohingya population in Myanmar, the role of speech and online propaganda following Russia's full-scale invasion of Ukraine, and the large-scale civilian harms associated with the ongoing conflict between Hamas and Israel. Taken together, these case studies demonstrate that the involvement of social media platforms is no longer peripheral but rather constitutes a recurring and structurally significant feature of contemporary atrocity contexts.

431. Building on an analysis of these conflicts, Chapter 3 examines the extent to which existing frameworks of international criminal law are capable of addressing the role of social media companies in facilitating or amplifying harmful speech. It concludes that, while corporate criminal liability for speech crimes is legally conceivable in principle, it remains largely unrealised in practice due to the relative absence of legal precedent and appropriate international fora at which such liability could be meaningfully exercised. The analysis therefore identifies a significant accountability gap at the international level; one that risks fostering an environment of impunity around powerful corporate actors operating in the digital sphere.

432. Looking beyond the corporate entity itself, Chapter 3 further explores the potential for individual criminal responsibility. It argues that senior executives and decision-makers within social media companies may, in certain circumstances, be subject to prosecution where harmful speech disseminated by their platforms satisfies the material and mental elements of international crimes, or where their conduct engages recognised modes of liability such as aiding and abetting or superior responsibility. In developing this argument, the chapter draws on lessons from domestic efforts to hold corporations and corporate actors accountable for speech-based harms, highlighting both the possibilities and the limitations of existing approaches.

433. Chapter 4 analyses the evidentiary framework governing the adjudication of speech crimes at the international level, with particular attention to content disseminated via social media platforms. It underscores that platforms differ significantly in their structural design, including the extent to which users may create groups and the degree of public accessibility to content shared within them. Admission mechanisms range from tightly moderated spaces controlled by human administrators to largely automated systems in which entry is granted upon completion of a standardised form. These structural distinctions may bear directly on legal characterisation, especially in relation to the 'public' element of the crime of direct and public incitement to genocide.

434. The analysis in Chapter 4 further highlights the forensic implications of platform-specific technical features. Variations in how platforms record timestamps, retain or strip metadata upon upload, and archive user activity can materially affect the evidentiary value of digital content. Metrics such as 'views' and 'impressions' are also examined as indicators of reach and potential impact, both of which are relevant when assessing the gravity and effect of harmful speech. In this context, the analysis revisits the role of AI-driven recommender systems and algorithmic amplification, noting that such systems may enhance the visibility of incendiary content and thereby expand its audience and influence.

435. Preservation of digital evidence emerges as a central concern. Social media companies play a critical role in retaining data that may be relevant to the investigation and prosecution of speech crimes and other atrocity offences. The chapter argues that where potentially incriminating material is removed from public view – whether pursuant to content moderation policies or otherwise – platforms should ensure its secure preservation for investigative purposes. In practice, however, content is often deleted or rendered inaccessible, complicating efforts by investigators and prosecutors to obtain reliable evidence.

436. In evaluating online speech, Chapter 4 emphasises the importance of contextual interpretation. Digital environments are characterised by diverse micro-cultures, each with distinct vocabularies, idioms, and communicative norms. Proper assessment of alleged speech crimes therefore requires sensitivity to these cultural and linguistic contexts in order to distinguish criminally relevant expression from speech that may be offensive or disturbing yet remains within the protective ambit of freedom of expression. The chapter refers in this regard to six indicia of speech crimes (particularly persecution) developed in investigations concerning Russian influencers in the context of the war in Ukraine.

437. The issue of verification is also addressed in light of growing concerns about manipulated and synthetic media. The increasing sophistication of deepfakes presents substantial evidentiary risks. Beyond reverse image searches, verification techniques include geo-location and chrono-location analysis, examination of the uploader's digital history, scrutiny of accompanying narration, identification of visual markers of authenticity, inspection of metadata, and corroboration through independent sources (triangulation). Such methods are essential to establishing authenticity and reliability in an environment where generative AI tools can produce convincing but fabricated content.

438. With respect to evidentiary presentation, Chapter 4 underscores the continuing relevance of the best evidence rule, which requires parties to submit the most reliable and complete version of a piece of evidence reasonably available in the circumstances. Although hearsay evidence is not per se inadmissible before international courts, chambers of the International Criminal Court have in recent years expressed caution regarding its weight and reliability. Given the prevalence of anonymous, unsourced, or AI-generated online materials, concerns surrounding hearsay are likely to intensify.

439. The chapter concludes with a cautionary note regarding procedural fairness. Effective access to technical expertise and investigative resources is indispensable for parties seeking to test the authenticity and reliability of digital evidence. A lack of transparency – whether in relation to platform algorithms, data retention practices, or evidentiary provenance – risks undermining the principle of equality of arms and impairing the ability of defendants to meaningfully challenge the evidence presented against them.

440. Chapter 5 examines the domestic prosecution of speech-related international crimes under the principle of universal jurisdiction. It argues that, as a matter of principle, speech may be prosecuted at the national level either as a standalone international crime or as a mode of liability, in a manner analogous to the practice of international courts and tribunals. Nonetheless, the chapter underscores that the practical realisation of such prosecutions presents a distinct set of legal and institutional challenges. Even where domestic legislation incorporates or mirrors international criminal law, the success of cases involving speech acts frequently depends on national investigative capacities, procedural frameworks, and doctrinal approaches.

441. The analysis identifies several points of divergence between domestic proceedings and the practice of international courts and tribunals, highlighting the particular difficulties associated with prosecuting speech through universal jurisdiction. Conversely, in certain respects, domestic systems may in fact provide a broader normative basis for accountability. For example, some national criminal codes contain incitement provisions extending beyond genocide to other core international crimes, and domestic aiding and abetting doctrines are not always constrained by the more restrictive 'purpose' requirement set out in Article 25(3)(c) of the Rome Statute of the International Criminal Court. These features may, in theory, facilitate a more expansive approach to liability for speech-based conduct via universal jurisdiction. International law is not the only game in town.

442. At the same time, universal jurisdiction cases raise acute evidentiary and contextual concerns. A central rationale of universal jurisdiction is to

enhance accountability for serious international crimes and gross human rights violations by permitting their prosecution in domestic courts irrespective of territorial or nationality links. However, in the specific context of speech crimes, national authorities operating at a geographical and cultural distance from the events in question may lack the linguistic fluency, socio-political understanding, or cultural competence necessary to accurately interpret the meaning and significance of contested statements. The risk is that speech may be overlooked, misconstrued, or undervalued as evidence of criminal conduct.

443. Unlike ad hoc or situation-specific international tribunals, which often benefit from dedicated resources – both human and financial – devoted to contextual and linguistic analysis throughout investigative and trial phases, domestic courts exercising universal jurisdiction may face more limited capacities. These constraints are particularly pronounced where the relevant languages are of limited diffusion or where online subcultures and coded forms of communication are involved.

444. Chapter 5 concludes by proposing that universal jurisdiction prosecutions of speech crimes should adhere to both minimum and aspirational standards – a 'floor and ceiling' approach. At a minimum, proceedings must comply with internationally recognised human rights guarantees, including the protection of freedom of expression and robust due process safeguards. At a higher level, domestic authorities should strive to implement best practices in evidence preservation, international cooperation, and the integration of advanced technological tools in the collection, analysis, verification, and presentation of digital speech evidence. Only through such safeguards can universal jurisdiction serve as a credible and principled mechanism for addressing speech-based international crimes.

445. Chapter 6 focuses on the most significant normative proposal advanced in these Guidelines: the call to amend Article 25(3)(e) of the Rome Statute. The current formulation and position of the article in the ICC Statute on direct and public incitement to genocide is unduly narrow and undermines the preventive function of international criminal law. As set out in detail in the chapter, there are normative, doctrinal, historical and moral reasons to support this amendment. It is time to expand the provision to cover incitement to all crimes within the Statute, while reaffirming the inchoate (illocutionary) nature of incitement to genocide and clarifying that liability does not depend on the subsequent commission or attempted commission of the incited crime. Incitement to the other crimes of war crimes, the crime of aggression, and crimes against humanity, does require evidence of commission and evidence to demonstrate the perlocutionary effects of a speech act

or acts. This statutory reform would better align the Rome Statute with contemporary risks and enhance the international community's ability to intervene before atrocities occur.

446. Having drawn together the key conclusions in each of the chapters, it is apposite to highlight overarching conclusions and cross-cutting points of concern. At the heart of these Guidelines is a structured method for evaluating speech acts. They propose a two-step process that begins with a contextual linguistic interpretation of the statements at issue, followed by a legal assessment of the extent to which those statements contributed to the commission of international speech crimes. This approach underscores the importance of situating speech within its social, political, and communicative environment, rather than treating it as an isolated utterance shorn of cultural or political context.

447. There are clear evidentiary challenges posed by digital communication. They call on governments and social media companies to create clear protocols to preserve potentially unlawful content and urge investigators to employ up-to-date technological tools to verify online material and document their investigative steps. In doing so, they acknowledge the fragility and manipulability of digital evidence and the need for rigorous methodological standards, as well as the need to protect privacy.

448. The dynamics of social media and how this impacts international criminal trials also require further attention. Our matrix for assessing the risk of a speech act highlights the significance of reach, amplification, and the speaker's authority in assessing the gravity and perlocutionary impact of speech acts. The architecture and affordances of digital platforms can dramatically magnify the effects of harmful speech, making traditional legal categories insufficient unless adapted to contemporary communicative realities.

449. The Guidelines also offer a nuanced conceptualisation of disinformation, recognising its capacity to facilitate crimes at multiple temporal stages: by inciting violence beforehand, enabling or legitimising violence during its commission, or obscuring responsibility afterward. They propose an inquiry into disinformation that encompasses knowledge of the underlying crimes, deliberate dissemination of falsehoods, and intent to further criminality, thereby providing a structured basis for legal evaluation.

450. Our analysis needs to extend beyond individual liability to also encompass universal jurisdiction and corporate responsibility. We recommend that domestic systems exercising universal jurisdiction incorporate these recommendations into national law while adhering to international human rights standards, particularly with respect to freedom of expression and due process. We advocate for the extension of corporate responsibility to speech crimes,

proposing a calibrated approach that reserves criminal liability for conduct meeting the gravity threshold while allowing for regulatory sanctions where the threshold is not met. Importantly, natural persons should not be able to evade accountability by acting through corporate structures.

451. The authors acknowledge that the present Guidelines are not intended to set out a comprehensive guide to all matters conceivably linked to speech crimes in the current era. Moreover, because they are legally oriented, they do not cover all potentially relevant disciplines. Despite not providing an exhaustive review of legal rules or procedures; rather, they function as a guide that can inform the development of more detailed regulatory instruments and practices. The rapidly evolving technological landscape, in which social media platforms, algorithmic amplification, and artificial intelligence increasingly shape the production, dissemination, and impact of speech globally call for new statutes and investigatory procedures. Nonetheless, the PJI Guidelines do not claim to have fully addressed the practical and jurisdictional challenges of enforcing speech-crime provisions across borders, especially in environments where digital platforms operate transnationally and state cooperation may be limited.

452. These gaps point toward several promising directions for future research and law reform. Further work is needed to clarify liability for AI-generated speech and to develop legal frameworks capable of addressing algorithmic amplification and synthetic media. Comparative research could illuminate how domestic free-speech doctrines intersect with international obligations, while interdisciplinary studies drawing on anthropology, psychology, and communication theory could deepen understanding of how speech mobilises collective violence. Additional refinement of the gravity threshold could assist courts and prosecutors in making consistent and principled assessments. Finally, the question of corporate accountability – particularly for social media platforms – requires sustained attention, as does the political and diplomatic strategy necessary to advance the proposed amendment to Article 25(3)(e). We conclude by urging States, civil society, and relevant institutions to work collaboratively toward implementing these recommendations and advancing the proposed amendment through the Assembly of States Parties.

Index

Ali Muhammed Ali Abd-al-Rahman (*The Prosecutor v Abd-al-Rahman*) 21–24
academic expression 2–3, 69–73, 125–137
accountability 1–3, 74–137, 153–180
actus reus 65
admissibility 138–152
Al Hassan Ag Abdoul Aziz Ag Mohamed Ag Mahmoud (*The Prosecutor v al-Hassan*) 63, 67
aiding and abetting 62–67, 94–116, 155–160
amplification 77–94, 99–105, 116–119, 141–147, 168–171, (algorithmic) 77–83, 89–94, 99–105
Artificial Intelligence (AI) generated content 89–94, 141–152, 173–180
Article 17(1)(d) Rome Statute 169
Article 19(3), ICCPR 135, 171
Article 25(3)(b) Rome Statute 18, 21, 63, 84, 165
Article 25(3)(c) Rome Statute 66–68, 113, 116, 175, 178
Article 25(3)(d) Rome Statute 56–67
Article 25(3)(e) Rome Statute 162–170
artificial intelligence (AI) 1–2, 68–70, 74–75, 89–94, 97–105, 137, 141–152, 168–180
asanacija 11
audience 3–5, 7–15, 77–94
authenticity 141–152

Belgium 158
best evidence rule 148, 178
Bosco Ntaganda (*The Prosecutor v Ntaganda*) 8–11, 18–21

Callixte Mbarushimana (*The Prosecutor v Mbarushimana*) 54–59
chain of custody xviii
Charles Blé Goudé (*The Prosecutor v Gbagbo and Blé Goudé*) 104
Charles Ghankay Taylor (*The Prosecutor v Taylor*) 108, 111, 113, 116
content moderation 97–115, 125–137
contextual analysis 4–7, 24–27, 31–33, 131–152
contribution to a crime 62–68, 94–116, 155–160
corporate responsibility 94–116, 169–171, 177–180
crime of aggression 163–168
crimes against humanity 28–73, 75–137, 153–180
cultural context 5–15, 24–27, 39–42, 50–54

deepfakes 89–94, 144–151
Democratic Republic of Congo (DRC) 8–10, 159–160
deterrence 18–22, 168–180
digital evidence 138–152, 173–180
direct and public incitement 64, 77–94, 105–108, 133–136, 155–168
directness 5–7, 24–27
disinformation 28–73, 74–94, 101, 125–126, 168–174
dissolution 169–171
domestic prosecution 119–124, 154–161
due process 12–15, 151–152, 160–167

Ellen Elias-Bursać 11–12
equality of arms 151–152, 178
erga omnes 153–154
Evariste Ntakirutimana 5
evidence preservation 141–147
evidence verification 144–147
executive liability 103–116
explicit speech 3–7

Facebook 30, 79–81, 101–102, 109–110, 112–113, 126–127, 131–132, 141
fair trial 12–15, 151–152
felicity conditions 4
Ferdinand Nahimana (*The Prosecutor v Nahimana et al.*) 3, 75–77, 89–94, 106
fines 168–171
Force v Facebook (United States) 101–102
freedom of expression 2–3, 29–31, 68–73, 89–94, 125–137, 161–167
freedom of information 125–137

© 2027 THE AUTHOR(S) | HTTPS://DOI.ORG/10.1163/9789004780262-011
This is an open access publication distributed under the terms of the CC 4.0 license BY-NC-ND.

General Comment No. 34 135
Genocide Convention 87, 163, 166
genocide denial 68–73
genocide 31–73, 77–94, 155–168
Germany 68–73, 158–160
Gonzalez v Google (United States) 101–102
gravity threshold 94–95, 122–125, 168–171

Hans Fritzsche (*Trial of the Major War Criminals before the International Military Tribunal*) 34–38
Hartford Guidelines 2, 6, 8, 76, 139, 141, 143–144, 158, 164, 168–169, 173
Hassan Ngeze (*The Prosecutor v Nahimana et al.*) 75–77, 106
hate speech 28–33, 68–73, 75–94, 125–133
hearsay 148–150, 178
Henry Kiprono Kosgey (*The Prosecutor v Ruto and Sang*) 59–62
human rights 2–3, 68–73, 89–94, 125–137, 151–152, 161–167

International Covenant on Civil and Political Rights (ICCPR) 33, 71, 136, 166
illocution 4–6
impact 3–5, 14–16, 77–94, 141–147, 168–171
implicit speech 3–7, 24–27
impunity 36–37, 64–73, 94–116, 153–180
incitement to genocide 3–5, 31, 63–64, 68–70, 77–94, 105–108, 119, 133–136, 155–168
information warfare 1–2, 74–94, 125–137
inkotanyi 6, 76, 133
instigation 16–27, 62–67, 94–116, 155–160
International Criminal Court (ICC) 1–3, 8–27, 54–70, 94–124, 138–152, 162–180
International Criminal Tribunal for Rwanda (ICTR) 3–15, 68–73, 75–94, 106
International Criminal Tribunal for the former Yugoslavia (ICTY) 5–15, 24–27, 34–54, 59–63, 68, 75–76
International human rights law 2–3, 68–73, 89–94, 125–137, 151–152, 161–167
International Military Tribunal (IMT) 34–39
investigative methods 139–152, 154–167
inyenzi 6
International Residual Mechanism for Criminal Tribunals (IRMCT) 24–27, 49–54
Israel 86–94

J.L. Austin 3–4
Jean-Bosco Barayagwiza (*The Prosecutor v Nahimana et al.*) 75–77, 104–107
Jean Paul Akayesu (*The Prosecutor v Akayesu*) 3–5
Jean Pierre Bemba Gombo (*The Prosecutor v Bemba*) 17, 63, 113, 147–150
joint criminal enterprise (JCE) 39–44, 62–67
Joshua Arap Sang (*The Prosecutor v Ruto and Sang*) 59–62
journalistic expression 2–3, 69–73, 125–137
Julius Streicher (*Trial of the Major War Criminals before the International Military Tribunal*) 38
jurisdiction 116–124, 153–167
jus cogens 153–154

Kass FM 59–62
key guidelines for the assessment of speech crimes under international criminal law 168–170
Kinyarwanda 5
Radislav Krstić (*The Prosecutor v Krstić*) 11
kupiga na kuchaji 8–10

licence restrictions 168–171
linguistic interpretation 7–15
locution 4–5

Mathias Ruzinda 5–6
mens rea 65–68
metadata 139–150
Milan Gvero (*The Prosecutor v Popović et al.*) 39–44, 62
misinformation 28–73, 75–94
modes of liability 16–27, 62–67, 94–116, 155–160
Murray v Feast (Australia) 14–16
Myanmar 78–81

national criminal law 154–167

New TV S.A.L. (*Prosecutor v New TV S.A.L. and Al Khayat*) 95–98
North Atlantic Treaty Organization (NATO) 40–43
natural persons 103–116

occupied Palestinian territory (oPt) 86–94
online speech 74–137
ordering 17–27

perlocution 4–5
platform liability 99–116
preventive function 163–167
probative value 147–152
propaganda 28–40, 44–54, 59–65, 75–94

Radio Télévision Libre des Mille Collines (RTLM) 6–8, 75–77
Radovan Karadžić (*The Prosecutor v Karadžić*) 24–25, 49–54, 59, 62–63
Ratko Mladić (*The Prosecutor v Mladić*) 11, 25–27, 44–49, 52, 59, 62–63
reach 20–22, 77–94, 100–105, 168–171
reasonable doubt 11–15, 68–70, 151–152
recommender systems 101–105
reliability 141–152, 177–178
retractions 169–171
risk assessment model 170–172
Rodrigo Roa Duterte (*The Prosecutor v Duterte*) 13
Rohingya 78–81, 86–94
Rome Statute of the International Criminal Court (Rome Statute) 3, 8, 16–27, 56–67, 82–87, 94–95, 109, 113–117, 122–124, 155–170, 177–180
Russia 81–86
Rwanda 3–15, 68–73, 75–94

Section 230, Communications Decency Act (United States) 100–102
shadow-banning 99–100
Sierra Leone 12–14
social media algorithms 99–105, 108–116
social media 74–137, 139–151, 168–171
socio-linguistic experts 14–16
speaker authority 20–22, 77–94, 103–116
Special Court for Sierra Leone (SCSL) 12–15
Special Tribunal for Lebanon (STL) 95–100
speech act theory 3–5
speech acts 1–27, 74–75, 138–152, 168–180
Srebrenica 11, 39–54, 62
state sovereignty 153–160

tampering 144–151
temporality 14–16, 28–30, 100–103, 141–147, 168–171
translation 7–15, 39–42, 152, 160
Twitter v Taamneh (United States) 101–103

Ukraine 81–86
universal jurisdiction 153–167, 177–179
United Nations Protection Force (UNPROFOR) 39–43

Vujadin Popović (*The Prosecutor v Popović et al.*) 11, 39–43, 62

war crimes 28–73, 75–137, 153–180
Western Australia 14–16
William Samoei Ruto (*The Prosecutor v Ruto and Sang*) 59–64
William Schabas 15

Zyklon B (*Case No. 9* (*The Zyklon B. Case: Trial of Bruno Tesch and Two Others*)) 107–108, 111–116

Speech has become an increasingly central feature of international criminal proceedings, particularly in cases involving senior political or military leaders. The PJI Guidelines are grounded in extensive jurisprudence, empirical research, and scholarly analysis, drawing on decisions from international courts and tribunals as well as broader interdisciplinary insights. Anchored on the fundamental right to freedom of expression, the Guidelines conveys guidelines for securing criminal accountability. In doing so, the book addresses cutting edge issues including disinformation, interpretation, social media, and AI propagated inflammatory and hateful speech, among other topics. It proposes 10 Key Guidelines, including an amendment to Article 25(3)(e) to criminalise intentional, direct and public incitement to any atrocity crime under the Rome Statute.

Contributors are: Katharina Braun, Juan Manuel Klein, Damla Yasemin Ipek, and Manuel J. Ventura.

Matthew Gillett, (Ph.D.), Leiden University, is Associate Professor of International Law at the University of Essex and Chair Rapporteur of the United Nations Working Group on Arbitrary Detention. He has won multiple awards for his research and writing.

Vanessa Topp is a Ph.D. candidate at the University of Essex Human Rights Centre. She has published several articles on international human rights law.

Richard Ashby Wilson is Professor of Anthropology and Co-Director of the Human Rights Initiative at Princeton University, and author of *Incitement on Trial: Prosecuting International Speech Crimes* (Cambridge University Press, 2017).

ISBN 978-9004-78025-5

9 789004 780255

brill.com