ACMUN 2026

BACKGROUND GUIDE

SOCIAL, HUMANITARIAN AND CULTURAL COMMITTEE (SOCHUM)

Agenda:

Combating Hate Speech and Disinformation while Upholding Freedom of Expression

LETTER FROM THE EXECUTIVE BOARD

Dear Delegates,

It is our privilege to welcome you to the Social, Humanitarian and Cultural Committee (SOCHUM) at this conference. The agenda before this committee sits at the intersection of some of the most fiercely contested values in contemporary international law: the right to speak freely, and the right to be protected from speech that incites hatred, violence, and discrimination, all while a third force — disinformation, industrialised and algorithmically amplified — corrodes the very possibility of informed public discourse.

Freedom of expression, protected under Article 19 of the International Covenant on Civil and Political Rights, is not an absolute right. It coexists in permanent tension with Article

20's prohibition on advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence. This tension is not a drafting oversight; it is a deliberate architecture that every state party has struggled to apply consistently. The emergence of social media platforms, algorithmic content amplification, generative artificial intelligence capable of producing synthetic text, images, and video, and coordinated state and non-state disinformation campaigns has placed extraordinary new pressure on that architecture. The legal instruments this committee will work with were drafted for a world of print, radio, and television. They were not drafted for a world in which a single fabricated video can reach hundreds of millions of people before any fact-checker has had the chance to respond.

This background guide is designed to be a neutral resource. It does not advocate for any particular regulatory model, any particular reading of contested legal thresholds, or any particular balance between expression and restriction. Where genuine legal, empirical, or political disagreement exists — and in this agenda, it exists almost everywhere — this guide presents the competing positions and the evidence marshalled in support of each, without endorsing one over another. It is the delegates' task, not the Executive Board's, to determine where that balance should lie, and to defend that position through the lens of the state they represent rather than through personal conviction.

This background guide does not provide you with settled answers. The line between protected offensive speech and prohibited incitement is genuinely contested, the technology of information dissemination is evolving faster than the law that governs it, and the geopolitical stakes are high, given that hate speech and disinformation regulation is also, in the hands of some governments, a documented tool of censorship. What this guide does is equip you with the legal foundations, technical context, and argumentative frameworks necessary to engage in substantive, rigorous debate. Your task as delegates is not to cite law for its own sake, but to use legal and technical reasoning to build real, implementable solutions to a real problem, while representing the authentic policy position of your assigned state.

With regards,

Priyanshu Biswas - Co Chairperson
Simra Fathima - Co Chairperson

INTRODUCTION TO THE COMMITTEE

A. Mandate and Institutional Position

The Social, Humanitarian and Cultural Committee is the Third Committee of the United Nations General Assembly. It is responsible for the examination of a wide range of social, humanitarian, and human rights questions affecting people worldwide, including questions relating to human rights, freedom of expression and the media, social development, crime prevention, refugees, and the advancement of women, children, and other vulnerable groups. SOCHUM operates under the broader framework of the United Nations Charter, drawing its authority from Articles 1(3), 13(1)(b), 55, and 62, which task the General Assembly and the Economic and Social Council with promoting universal respect for, and observance of, human rights and fundamental freedoms for all, without distinction as to race, sex, language, or religion.

Like all six Main Committees of the General Assembly, SOCHUM's output takes the form of draft resolutions, which are debated and voted upon in committee before being forwarded to the GA plenary for final adoption. SOCHUM resolutions are not legally binding in the manner of Security Council resolutions adopted under Chapter VII of the Charter; the General Assembly is fundamentally a deliberative and recommendatory body under Article 10 of the Charter. However, SOCHUM resolutions carry substantial normative weight. They have historically shaped the drafting of binding human rights treaties (the Universal Declaration of Human Rights itself began as a Third Committee product before its adoption by the plenary in 1948), the mandates of UN human rights mechanisms such as Special Rapporteurs, and the interpretive practice of treaty bodies including the Human Rights Committee. A SOCHUM resolution on this agenda would not create binding law, but it could establish the terms of reference for future treaty negotiation, direct UN agencies toward specific programmatic priorities, and articulate an authoritative — if non-binding — statement of the international community's evolving position.

B. Relationship to Other UN Bodies

Delegates should understand that SOCHUM does not operate in isolation. The Human Rights Council (a subsidiary body of the General Assembly, though institutionally distinct from SOCHUM) maintains the mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, whose reports form a substantial part of the soft-law foundation for this agenda. The Human Rights Committee, established under the ICCPR's First Optional Protocol, issues General Comments interpreting Covenant obligations and considers individual communications alleging violations. UNESCO, a specialised agency rather than a General Assembly subsidiary body, has developed extensive technical guidance on media literacy, platform governance, and the safety of journalists. The Security Council has, in narrow circumstances (such as its consideration of Da'esh/ISIL propaganda under Resolution 2354 (2017)), addressed incitement to terrorism as a threat to international peace and security. This committee should be conscious of this broader institutional ecosystem when drafting resolution clauses — recommending action, requesting reports, or establishing mechanisms in ways that are coherent with, rather than duplicative of, the existing architecture.

C. Why This Agenda Belongs in SOCHUM

The committee's relevance to hate speech and disinformation is both immediate and structural. Neither phenomenon fits cleanly within a single existing regulatory regime. Hate speech implicates the ICCPR, the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), international humanitarian law and international criminal law in situations of armed conflict or atrocity, and a patchwork of national criminal codes with widely divergent thresholds. Disinformation implicates none of these directly, since no binding international instrument defines or prohibits "disinformation" as such; it instead falls into gaps between electoral law, data protection law, platform governance, consumer protection law, and freedom of expression jurisprudence. SOCHUM is precisely the forum where the fragmentation of these regimes can be identified, debated, and addressed through the development of new normative frameworks, while remaining acutely conscious that any expansion of restrictions on expression carries a documented risk of being turned against journalists, human rights defenders, political opposition figures, and minority communities.

HISTORICAL AND CONTEXTUAL BACKGROUND

A. Propaganda in the Pre-Digital Era

State and non-state use of mass communication to incite hatred and violence has a long and deeply troubling history that long predates the internet, and understanding this history is essential to understanding why the international community has treated incitement as a distinct category of harm requiring legal prohibition rather than merely social disapproval.

The twentieth century witnessed the systematic use of radio and print propaganda to prepare populations for mass atrocity. Nazi Germany's state and party-affiliated press, most infamously the newspaper *Der Stürmer*, combined with state radio broadcasting under Joseph Goebbels' Reich Ministry of Public Enlightenment and Propaganda, engaged in a sustained campaign of dehumanising rhetoric against Jewish, Roma, disabled, and other targeted populations over more than a decade, a campaign the Nuremberg Tribunal and subsequent historical scholarship have identified as a necessary precondition for the social acceptability of the Holocaust.[1] The trial of publisher Julius Streicher at Nuremberg — convicted specifically for incitement to murder and extermination through his editorial control of *Der Stürmer*, despite never personally participating in any killing — established the foundational international law principle that speech alone, absent any direct physical participation in violence, can constitute a prosecutable international crime where it incites atrocity at scale.[2]

The 1994 Rwandan genocide provides the most extensively studied and legally consequential case study in this domain. Radio Télévision Libre des Mille Collines (RTLM), a private radio station with close links to the Hutu Power movement, broadcast a sustained campaign from 1993 through 1994 that explicitly and repeatedly referred to Tutsi civilians as "inyenzi" (cockroaches), called for their extermination, and in some documented instances broadcast the specific locations of individuals in hiding, directly facilitating their murder. The International Criminal Tribunal for Rwanda's landmark "Media Case" judgment against RTLM founder Ferdinand Nahimana, *Kangura* newspaper editor Hassan Ngeze, and RTLM chairman Jean-Bosco Barayagwiza found each guilty of direct and public incitement to genocide, conspiracy to commit genocide, and genocide itself, establishing the most authoritative judicial precedent in international law for the proposition that mass media incitement can constitute genocide even where the disseminator does not personally wield a weapon.[3] The ICTR's reasoning in this case remains the single most frequently cited judicial authority in contemporary debates about the incitement threshold, and delegates researching the legal architecture of this agenda are strongly encouraged to read the judgment directly.

B. The Rise of the Internet and Web 2.0

The migration of public discourse onto the internet in the late 1990s and 2000s initially inspired significant optimism among human rights advocates, who viewed the technology's decentralising and disintermediating character as inherently favourable to freedom of expression, since it reduced dependence on state-controlled or commercially gatekept broadcast media. Early internet governance debates, including those surrounding the 2003 and 2005 World Summits on the Information Society, largely proceeded from this optimistic premise, emphasising universal access and the internet's democratising potential.

This optimism was not unfounded with respect to certain dimensions of the technology — the internet did meaningfully lower the cost of independent journalism, cross-border civil society organising, and minority-language publishing. However, the same low barrier to entry that enabled this democratisation also enabled the proliferation of extremist forums, coordinated harassment campaigns, and early large-scale disinformation operations, particularly around the mid-2000s emergence of social networking platforms (Friendster, MySpace, and subsequently Facebook and Twitter). The technical and institutional infrastructure for content moderation at scale — the systems of policy-writing, human review, and automated detection that platforms rely on today — did not yet exist in any mature form during this period, meaning platforms largely responded to harmful content reactively, on a complaint-driven basis, rather than through systematic prevention.

C. Platformisation and the Attention Economy

The dominant transformation of the following decade was what scholars have termed the "platformisation" of public discourse: the migration of political speech, news consumption, and interpersonal communication onto a small number of dominant social media platforms — Facebook (subsequently rebranded Meta), Twitter (subsequently rebranded X), YouTube, Instagram, TikTok, and WhatsApp — which concentrated unprecedented power over the visibility of speech in the hands of private corporations operating largely outside binding international human rights obligations, and whose business models depend overwhelmingly on advertising revenue tied to user engagement and attention.

Engagement-optimised recommendation algorithms did not merely provide a new venue for existing hate speech and disinformation; they created qualitatively new incentive structures that reward emotionally provocative, polarising, and often false content with disproportionately greater reach compared to neutral or nuanced content. This dynamic was documented extensively in leaked internal research from Meta disclosed by whistleblower Frances Haugen in 2021 (the "Facebook Papers"), which revealed that company researchers had internally identified the amplifying effect of the platform's ranking algorithm on divisive and anger-inducing content years before the disclosure became public, and that internal proposals to reduce this amplification had been shelved due to their projected negative effect on engagement metrics and, consequently, advertising revenue.[4]

D. The 2016 Inflection Point: Brexit, the US Election, and Cambridge Analytica

The 2016 Brexit referendum in the United Kingdom and the 2016 United States presidential election marked a widely recognised inflection point in public and governmental understanding of digital influence operations. Subsequent investigations — including the UK House of Commons Digital, Culture, Media and Sport Committee's extensive "Disinformation and 'fake news'" inquiry, which produced its final report in 2019 — documented the use of microtargeted political advertising, coordinated inauthentic account networks, and improperly harvested personal data to influence electoral outcomes at a scale and level of sophistication not previously understood by regulators or the public.[5] The Cambridge Analytica scandal, in which a political consulting firm obtained the personal data of an estimated 87 million Facebook users without meaningful consent and used psychographic profiling techniques to microtarget political advertising, became the most publicly recognised symbol of this transformation, prompting the resignation of Cambridge Analytica's leadership, substantial regulatory fines against Facebook in multiple jurisdictions, and a wave of national parliamentary inquiries into platform accountability.

The US Department of Justice's investigation into Russian interference in the 2016 election (the "Mueller Report," 2019) documented, with formal governmental attribution, a coordinated operation by Russia's Internet Research Agency involving thousands of social media accounts designed to sow social and political division within the American electorate, alongside a separate hacking-and-leaking operation targeting the Democratic National Committee and the Clinton campaign.[6] This episode remains one of the most extensively documented and formally attributed state-sponsored disinformation operations in the historical record, and is frequently invoked in this committee's debates by delegations on multiple sides of the sovereignty-versus-accountability divide.

E. The Rohingya Crisis: Genocide Amplification at Platform Scale

The clearest and most devastating illustration of the stakes involved in this agenda is the role of Facebook in the 2016–2017 violence against the Rohingya Muslim population in Myanmar's Rakhine State. An independent Human Rights Impact Assessment commissioned by Meta itself and conducted by the non-profit organisation BSR concluded that Facebook had become a platform used to "foment division and incite offline violence" against the Rohingya, and that the company had not done enough, quickly enough, to prevent its platform from being used to that effect, notwithstanding repeated warnings from Myanmar-based civil society organisations beginning as early as 2013.[7]

United Nations investigators reached similar conclusions through an independent process. The UN Independent International Fact-Finding Mission on Myanmar found that Facebook had played a "determining role" in stoking the hatred that preceded the violence against the Rohingya, and characterised the broader campaign of violence as bearing the hallmarks of genocide, citing the systematic dehumanisation of the Rohingya population across Facebook — then the dominant, and for many users the only, means of accessing the internet in Myanmar, following the country's rapid post-2011 liberalisation of telecommunications markets.[8] Reporting by Reuters and other outlets subsequently documented that Meta had, at the height of the crisis, employed only a handful of Burmese-language content moderators to review content generated by millions of Myanmar-based users, illustrating a resourcing and linguistic-capacity gap that remains a recurring theme throughout this agenda: platforms' content moderation infrastructure is disproportionately concentrated in English and a small number of other high-resource languages, leaving populations who communicate in lower-resource languages significantly more exposed to unmoderated hate speech and disinformation.

F. The COVID-19 "Infodemic"

The COVID-19 pandemic, beginning in early 2020, produced what the World Health Organization formally termed an "infodemic" — an overabundance of information, including deliberately false and misleading content, that made it difficult for people to find trustworthy sources and reliable guidance when they needed it, with direct and measurable public health consequences.[9] Documented pandemic-era disinformation included false claims about the origins of the virus, fabricated and sometimes lethal "miracle cure" claims (including a documented spike in methanol poisoning deaths in Iran linked to false claims that consuming alcohol prevented infection), and a sustained global anti-vaccination disinformation campaign that multiple studies subsequently associated with reduced vaccine uptake and, in specific documented instances, preventable deaths.

The pandemic period is significant for this committee's purposes because it produced an unusually rapid and coordinated — though also unusually controversial — platform and state response, including expanded platform content-labelling systems, WHO-platform collaboration on authoritative information promotion, and, in a substantial number of states, new or expanded "fake news" criminal statutes specifically targeting pandemic-related disinformation. Civil society organisations and the UN Special Rapporteur on freedom of expression subsequently documented that a meaningful proportion of these pandemic-era "fake news" laws were applied not against genuine disinformation actors but against journalists and civil society figures critical of government pandemic response, illustrating in real time the weaponisation risk that runs throughout this agenda.[10]

G. The Generative AI Era

The most recent transformation shaping this agenda is the rapid public availability, beginning around 2022–2023, of large language models capable of generating fluent, persuasive text at scale, and of diffusion-model-based systems capable of generating photorealistic synthetic images, audio, and video — commonly termed "deepfakes." The detection in January 2024 of AI-generated robocalls impersonating the voice of US President Joe Biden, deployed to discourage voter turnout ahead of the New Hampshire Democratic primary, demonstrated in a live electoral context that synthetic media generation had moved decisively from theoretical risk to operational tool, prompting a formal declaratory ruling from the US Federal Communications Commission treating AI-generated voice calls as subject to existing robocall regulation.[11] Numerous national elections held in 2024, described by some commentators as the "largest election year in human history," were accompanied by documented instances of AI-generated disinformation, though independent post-election analyses in several cases found that the overall electoral impact of such content remained difficult to establish with confidence — a methodological difficulty that recurs throughout the empirical literature on disinformation generally, and which delegates should bear in mind: causal claims about disinformation's electoral or social impact are frequently asserted with more confidence than the underlying evidence supports.

---

[1] Raul Hilberg, *The Destruction of the European Jews* (1961; rev. ed. 2003), Yale University Press. [2] International Military Tribunal, *Trial of the Major War Criminals*, Judgment (1946), Nuremberg (conviction of Julius Streicher). [3] *Prosecutor v. Ferdinand Nahimana, Jean-Bosco Barayagwiza and Hassan Ngeze* ("Media Case"), ICTR-99-52-T, Judgment (3 December 2003). [4] Jeff Horwitz, "The Facebook Files," *The Wall Street Journal* (2021 investigative series); US Securities and Exchange Commission, Whistleblower Complaint of Frances Haugen (2021). [5] UK House of Commons, Digital, Culture, Media and Sport Committee, *Disinformation and 'fake news': Final Report* (2019), HC 1791. [6] US Department of Justice, *Report on the Investigation into Russian*

Interference in the 2016 Presidential Election ("Mueller Report") (2019), Vol. I. [7] BSR, *Human Rights Impact Assessment: Facebook in Myanmar* (2018). [8] UN Human Rights Council, *Report of the Independent International Fact-Finding Mission on Myanmar*, A/HRC/39/64 (2018), paras 74–76. [9] World Health Organization, *Infodemic Management: An Overview of Infodemiology* (2020). [10] UN Human Rights Council, Report of the Special Rapporteur on freedom of opinion and expression, A/HRC/47/25 (2021). [11] Federal Communications Commission, Declaratory Ruling on AI-Generated Voice Robocalls (FCC 24-16, 2024).

CONCEPTUALISING HATE SPEECH AND DISINFORMATION

A. Definitional Challenges

The most fundamental gap in the existing regulatory framework is definitional. There is no single, internationally agreed, legally binding definition of "hate speech," and none whatsoever of "disinformation." This absence is not merely a technical inconvenience; it is a structural vulnerability. International law generally cannot prohibit, restrict, or regulate what it cannot define with reasonable precision, particularly where the object of regulation is speech itself, given the well-documented chilling effect that vague prohibitions have on legitimate expression, journalism, and dissent.

The United Nations Strategy and Plan of Action on Hate Speech (2019) offers a working, non-legally-binding definition: hate speech is "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."[12] This definition is deliberately broad, precisely because the UN Strategy is a policy and programmatic instrument rather than a criminal prohibition. This reveals the core definitional dilemma facing the committee: a definition broad enough to capture the phenomenon delegates wish to address is almost always broad enough to capture protected speech as well, including satire, academic critique of religious doctrine, and robust political criticism.

Disinformation presents an even more acute definitional problem, because — unlike hate speech, which can in principle be identified by its content and target — disinformation is defined by the falsity and intent of the speaker, both of which are extraordinarily difficult to establish through law and to prove through evidence admissible in a legal proceeding. UNESCO's guidance for policymakers distinguishes three related but legally distinct phenomena: "disinformation" (false information deliberately created and shared to cause harm), "misinformation" (false information shared without intent to cause harm, such as an individual sincerely but mistakenly sharing an inaccurate news article), and "malinformation" (genuine, accurate information shared with intent to cause harm, such as the non-consensual publication of leaked private data or "doxxing").[13] This tripartite distinction matters enormously for this committee, because a resolution that fails to distinguish intentional falsification from honest error, satire, or genuine opinion risks becoming a tool for suppressing legitimate journalism and dissent — precisely the abuse pattern already documented in numerous "fake news" laws enacted by governments across a wide range of political systems since 2018.

B. The Elements of Hate Speech: Content, Target, and Context

Scholarly and legal analysis generally identifies three necessary elements in assessing whether a given communication constitutes hate speech in the legally relevant sense. First, the **content element**: the communication must contain language that is pejorative, dehumanising, or that advocates hostility, discrimination, or violence. Second, the **target element**: the communication must be directed at a person or group identified by reference to a protected characteristic — race, ethnicity, national origin, religion, and, in an increasing number of national and regional frameworks, gender, sexual orientation, disability, or caste. Third, the **contextual element**: the same words may carry dramatically different legal and social significance depending on the speaker's identity and authority, the audience, the medium, the historical and political context, and the prevailing risk of violence — the same insulting remark directed at a majority group in a stable, low-tension society carries a different risk profile than an identical remark directed at a historically persecuted minority in a context of rising intercommunal violence.

C. The Elements of Disinformation: Falsity, Intent, and Harm

A parallel three-element framework is generally applied to disinformation. First, the **falsity element**: the claim must be objectively false or materially misleading, a standard that itself excludes matters of genuine opinion, prediction, satire, and areas of authentic scientific or scholarly uncertainty. Second, the **intent element**: the disseminator must know, or be reckless as to, the falsity of the claim, and must intend to deceive — this element is what distinguishes disinformation from the far larger category of misinformation, where a person sincerely but mistakenly shares false content believing it to be true. Third, the **harm element**: many proposed frameworks require that the false claim be capable of causing, or be intended to cause, public or individual harm — to public health, to electoral integrity, to public order, or to an individual's reputation or safety — rather than merely being false in a trivial or inconsequential sense.

For the purposes of this committee's work, delegates should understand hate speech to encompass expression that attacks, demeans, or incites hostility, discrimination, or violence against a person or group on the basis of a protected characteristic, and disinformation to encompass the deliberate creation and dissemination of false or materially misleading information, through text, image, audio, or video, with the intent to deceive and cause public or individual harm. Both working definitions are necessarily provisional and contested; their elaboration is itself a central task for this committee, and delegates are encouraged to propose refinements to these working definitions in their draft resolutions rather than treating them as fixed.

D. Why the Two Phenomena Are Treated Together

Hate speech and disinformation are analytically distinct — one is defined primarily by its target and content, the other by its truth-value and intent — yet this agenda treats them together for good reason. Empirically, the two phenomena are frequently co-occurring and mutually reinforcing: disinformation campaigns frequently deploy hateful and dehumanising framing of a targeted group as a mobilisation and persuasion technique (as documented extensively in both the Rwandan and Myanmar case studies above), while hate speech campaigns frequently rely on fabricated atrocity claims, false statistics, or manufactured incidents to justify hostility toward a target group. The two phenomena also share a common institutional battleground: both are primarily disseminated today through the same small set of dominant social media platforms, are subject to the same recommendation-algorithm amplification dynamics, and raise structurally identical tensions between restriction and protected expression under Article 19 of the ICCPR.

---

[12] United Nations, *United Nations Strategy and Plan of Action on Hate Speech* (2019). [13] UNESCO, *Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training* (Ireton & Posetti eds., 2018).

CORE COMPONENTS AND CATEGORIES

Hate speech and disinformation, though distinct, can be disaggregated into interoperable components relevant to regulation. The **content component** encompasses the actual claim, image, or statement — the classical subject matter of both defamation law and incitement law. The **amplification component** comprises the algorithmic recommendation systems, paid advertising infrastructure, and automated bot networks that determine reach and virality, largely outside any binding transparency obligation in most jurisdictions. The **coordination component** refers to the organisational infrastructure — troll farms, paid influence-for-hire networks, and state media apparatuses — that enables campaigns to operate at scale and across jurisdictions. The **platform governance component**, perhaps the most institutionally significant, describes the content moderation policies, enforcement mechanisms, and appeals processes through which private companies — not states — currently make the overwhelming majority of real-world decisions about what speech is permitted to remain online.

Within this general framework, several distinct categories of harmful content warrant particular attention from delegates drafting resolution clauses.

Direct incitement content explicitly calls for violence, discrimination, or hostility against an identifiable group and sits closest to the settled prohibitions of Article 20(2) ICCPR; this is the category with the clearest existing legal basis for restriction.

Dehumanising and eliminationist content does not necessarily contain an explicit call to violence but systematically frames a target group in terms historically associated with violence against them — as vermin, disease, invaders, or subhuman — and occupies a harder legal category, since it may fall short of direct "incitement" in the strict Rabat Plan sense while still functioning, as the Rwandan and Myanmar case studies demonstrate, as a documented precursor to atrocity.

Coordinated disinformation campaigns use networks of authentic-seeming or automated accounts to manufacture the false appearance of grassroots consensus around a claim, a technique frequently termed "astroturfing."

Synthetic media ("deepfakes") use generative AI to fabricate audio, image, or video evidence of statements or events that never occurred, or to alter genuine media in materially misleading ways.

Algorithmic amplification refers not to any single piece of content but to the systemic tendency of recommendation systems to preferentially surface emotionally activating material, meaning that even lawful individual pieces of content can, in aggregate and at scale, produce harms that no single piece of content could produce alone — a distinctly modern regulatory problem with no clear analogue in pre-digital law.

State-sponsored information operations involve the direct or indirect use of state resources, whether through official state media, intelligence-linked front organisations, or paid proxies, to conduct coordinated hate speech or disinformation campaigns, often but not always targeting foreign populations, and raising distinct questions of state responsibility under international law that are addressed later in this guide.

Non-state and commercially motivated disinformation encompasses so-called "disinformation-for-hire" firms, financially motivated "clickbait" operations that profit from advertising revenue generated by viral false content regardless of political intent, and individually motivated harassment campaigns, all of which fall outside the state-attribution framework applicable to the category above.

SCIENTIFIC AND TECHNICAL FOUNDATIONS

A. Recommendation Algorithms and Engagement Optimisation

Modern social media platforms rely on machine-learning recommendation systems trained to maximise user engagement — time spent, clicks, shares, and comments — as the primary signal used to rank and surface content, because engagement is closely correlated with the advertising revenue that constitutes the overwhelming majority of these companies' income. Because emotionally intense content (anger, fear, moral outrage, in-group solidarity against an out-group) reliably generates higher engagement than neutral or nuanced content, these systems produce a structural, and in many documented cases largely unintended, bias toward amplifying precisely the categories of content this committee is concerned with.

Internal Meta research disclosed by whistleblower Frances Haugen found that the company's own data scientists had identified this amplification dynamic internally, and that a 2018 change to the platform's ranking algorithm intended to promote "meaningful social interactions" had the empirically documented side effect of rewarding outrage-inducing and divisive content with disproportionate reach, prompting some European political parties to report altering their own communications strategy toward more provocative content specifically in response to the algorithm's incentive structure.[14] This is the central regulatory implication for this committee: harmful outcomes at platform scale are frequently not the product of any individual user's or publisher's malicious intent but an emergent property of profit-optimised system design, meaning that content-level regulation alone — addressing individual posts after the fact — may be structurally insufficient without also addressing the amplification architecture itself, though delegates should note this is a contested empirical and policy claim, not a settled consensus, as discussed further below.

B. Filter Bubbles, Echo Chambers, and the Empirical Debate

A substantial body of academic literature has debated the extent to which algorithmic curation produces "filter bubbles" (a term coined by internet activist Eli Pariser) or "echo chambers" that systematically narrow individuals' exposure to diverse or challenging viewpoints, thereby increasing susceptibility to hate speech and disinformation within an ideologically homogeneous information environment. It is important for delegates to understand that this is a genuinely contested empirical question, not settled fact: while some studies find measurable ideological segregation in social media consumption patterns, other rigorous empirical studies — including large-scale randomised field experiments conducted in partnership with Meta and published in *Science* and *Nature* in 2023 — have found that algorithmic feed changes (including reverting users to purely chronological, non-algorithmic feeds) produced smaller effects on political polarisation and belief in false claims than the "filter bubble" hypothesis would predict, suggesting that offline social and political factors may play a larger causal role than algorithmic curation alone.[15] Delegates should treat causal claims about algorithmic amplification's real-world impact with appropriate empirical caution and should be prepared to cite specific studies rather than asserting the effect as self-evident.

C. Generative AI and Synthetic Media Detection

The application of generative adversarial networks (GANs) and, more recently, diffusion models and large language models to media synthesis has progressed from a research curiosity to a widely and cheaply accessible public capability within a remarkably compressed timeframe — roughly 2018 to the present. Deepfake detection technology has struggled to keep pace with generation technology: researchers at institutions including MIT's Media Lab have repeatedly demonstrated that detection accuracy degrades rapidly as generation techniques improve, creating an adversarial "arms race" dynamic structurally similar to that observed in cybersecurity, where each advance in detection prompts a corresponding advance in evasion.[16]

The critical regulatory implication is that any framework relying primarily on after-the-fact technical detection as a governance mechanism will likely always lag behind the generative capability it seeks to police. This has led a substantial number of technologists and policymakers to argue that governance frameworks should focus at least as much on **provenance** — verifying and labelling the origin and authenticity of genuine media at the point of creation, for example through the Coalition for Content Provenance and Authenticity's (C2PA) cryptographic content-authenticity standard — as on the comparatively harder problem of detecting fabrication after publication.

D. Content Moderation at Scale: Human Review and Automated Systems

Content moderation at the scale operated by major platforms — Meta's platforms alone process content from more than three billion monthly active users — is not, and cannot practically be, conducted entirely by human reviewers. Platforms rely on a hybrid system combining automated classifiers (machine-learning models trained to detect specific categories of prohibited content, such as terrorist content, child sexual abuse material, or graphic violence, with generally high accuracy for well-defined visual categories) and human review teams (frequently third-party contracted workforces based in lower-wage countries, whose working conditions and psychological exposure to traumatic content have themselves become the subject of significant labour-rights litigation and reporting).[17]

Automated classifiers perform substantially less reliably on hate speech and disinformation than on more visually or structurally distinctive content categories, because both phenomena depend heavily on linguistic and cultural context, sarcasm, coded language ("dog whistles"), and rapidly evolving slang that automated systems, particularly those trained predominantly on English-language data, frequently fail to detect accurately, or conversely flag over-broadly in ways that suppress legitimate speech such as counter-speech, satire, or academic discussion of hateful ideologies. This technical limitation is directly relevant to the linguistic-resourcing gap discussed in the Rohingya case study above and returns repeatedly throughout this guide's discussion of regulatory gaps.

E. Bot Networks and Coordinated Inauthentic Behaviour Detection

Platform-side detection of coordinated inauthentic behaviour relies on network analysis identifying clusters of accounts exhibiting synchronised posting patterns, shared technical infrastructure (common IP address ranges, device fingerprints, or account creation patterns), and content duplication inconsistent with organic, independent human behaviour. The Oxford Internet Institute's Computational Propaganda Project has documented organised social media manipulation campaigns across dozens of countries, finding that state actors, political parties, and private "disinformation-for-hire" contractors all participate in this activity, and that the sophistication and commercial availability of such campaigns has increased substantially even as public and regulatory awareness of the phenomenon has grown.[18] Meta's own quarterly Adversarial Threat Reports have documented and removed hundreds of distinct state-linked influence operations originating from a broad range of countries across every world region since 2017, illustrating that such campaigns are not a fringe or geographically isolated phenomenon but a persistent, well-resourced, and geographically dispersed feature of the contemporary global information environment.

---

[14] Jeff Horwitz, "The Facebook Files," *The Wall Street Journal* (2021). [15] Guess et al., "How do social media feed algorithms affect attitudes and behavior in an election campaign?", *Science* 381 (2023); Nyhan et al., "Like-minded sources on Facebook are prevalent but not polarizing," *Nature* 620 (2023). [16] MIT Media Lab, Detect Fakes Project, ongoing research publications (2020–present). [17] Casey Newton, "The Trauma Floor," *The Verge* (2019) (investigative reporting on content moderator working conditions). [18] Oxford Internet Institute, *Industrialized Disinformation: 2020 Global Inventory of Organised Social Media Manipulation* (2021).

EXISTING INTERNATIONAL LEGAL FRAMEWORK

The legal landscape governing hate speech and disinformation is defined not by a single coherent regime but by a constellation of partially applicable instruments whose collective coverage of the threat is deeply inadequate — and, in several important respects, is inadequate by deliberate design, since excessive restriction of expression is itself among the harms this framework exists to prevent.

A. Article 19 of the ICCPR: The Baseline Right

Article 19 of the International Covenant on Civil and Political Rights (1966), ratified by 174 states as of the most recent depositary records, establishes the core protection: everyone has the right to freedom of expression, including the freedom to seek, receive, and impart information and ideas of all kinds, regardless of frontiers, through any media of the person's choice. This right is not absolute. Article 19(3) permits restrictions, but only where they satisfy a cumulative three-part test: restrictions must be (1) **provided by law** (meaning formulated with sufficient precision to allow an individual to regulate their conduct accordingly, rather than left to unfettered administrative discretion), (2) **necessary** (meaning the restriction must be the least intrusive instrument among those which might achieve the desired protective result — a genuine necessity, not merely a convenient or reasonable one), and (3) directed toward a **legitimate aim**, namely respect for the rights or reputations of others, or the protection of national security, public order (*ordre public*), public health, or public morals.[19] This tripartite legality-necessity-legitimacy test is the foundational analytical framework this committee must apply to any proposed restriction, and delegates are expected to test every regulatory proposal discussed in this committee against it.

The Human Rights Committee's General Comment No. 34 on Article 19 (2011) provides the most authoritative interpretive guidance available. It emphasises that restrictions must be proportionate to the specific interest they protect and must not put the right itself in jeopardy, and that the value placed by the Covenant upon uninhibited expression is particularly high in circumstances of public debate concerning public figures in the political domain and public institutions.[20] General Comment No. 34 explicitly cautions against broadly framed "extremism," "blasphemy," or "memory law" restrictions, noting the documented risk that such laws are frequently applied in a manner incompatible with Article 19 to suppress legitimate criticism of governments, religious institutions, or historical narratives.

B. Article 20 of the ICCPR: The Incitement Threshold

Article 20(2) imposes the countervailing binding obligation most directly relevant to hate speech: "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law." Unlike Article 19(3), which merely *permits* restriction at a state's discretion, Article 20(2) affirmatively *requires* it — states parties are under a positive obligation to enact domestic legislation prohibiting this specific category of expression. Notably, several states — including the United States, on ratification of the ICCPR in 1992 — entered formal reservations to Article 20, on the basis that domestic constitutional free-speech protections were considered incompatible with a mandatory prohibition of this kind, illustrating from the outset the tension this committee is tasked with addressing.

The critical interpretive difficulty — and the central legal question this committee must grapple with — is the threshold between mere advocacy of hatred (which may be deeply offensive but is not, by itself, an internationally mandated basis for prohibition) and advocacy that rises to the level of constituting incitement (which is mandatorily prohibited). The Covenant's text does not itself define this threshold with precision, and state practice in operationalising it varies enormously.

The **Rabat Plan of Action**, adopted following a series of OHCHR-convened expert workshops held across four regions in 2011, remains the most authoritative attempt to operationalise this threshold. It proposes a six-part contextual test for assessing whether expression rises to the level of prohibited incitement: (1) the broader **social and political context** in which the statement was made and disseminated; (2) the **status of the speaker**, including their position of authority or influence over the relevant audience; (3) the demonstrable **intent** to incite the audience against a target group, as distinct from merely provocative or offensive intent; (4) the **content and form** of the speech itself, including its degree of provocation and directness; (5) the **extent of its dissemination**, including the size, means, and reach of its audience; and (6) the **likelihood of harm**, including a reasonable probability that the speech would succeed in inciting actual action against the target group, and the imminence of that risk.[21] The Rabat threshold test is not itself a legally binding instrument, but it has been widely cited by the Human Rights Committee, successive Special Rapporteurs on freedom of opinion and expression, and a number of domestic and regional courts as the leading interpretive framework for applying Article 20(2), and delegates are strongly encouraged to consider building resolution language around this six-part test, whether adopting it, refining it, or explicitly rejecting elements of it.

C. The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)

ICERD, adopted in 1965 and one of the most widely ratified human rights treaties with 182 states parties, imposes obligations that in some respects go further than the ICCPR. Article 4 requires states parties to declare an offence punishable by law "all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination," as well as the financing and organisation of such activities, and to declare illegal and prohibit organisations which promote and incite racial discrimination.[22] The Committee on the Elimination of Racial Discrimination's General Recommendation No. 35 (2013) provides interpretive guidance on Article 4, emphasising a similar contextual and proportionality-based approach to that found in the ICCPR framework, while maintaining that the provision's textual scope is, on its face, broader than Article 20(2) ICCPR.

Several Western states — including the United States, the United Kingdom (historically), and others — entered reservations to Article 4 on the basis that its literal text is difficult to reconcile with robust free expression protections, since it appears on its face to require prohibition of mere "dissemination of ideas based on racial superiority" without the additional incitement or imminence threshold found in Article 20(2) ICCPR. This divergence between ICERD's textual breadth and the ICCPR's more protective architecture is one of the clearest and most frequently debated points of doctrinal tension in the existing framework, and a productive area for committee debate regarding how any new instrument should reconcile the two standards.

D. International Humanitarian Law and International Criminal Law

Where hate speech occurs in the context of armed conflict or contributes to the commission of core international crimes, a distinct body of law applies. The Rome Statute of the International Criminal Court (1998), Article 25(3)(e), establishes individual criminal responsibility for a person who "directly and publicly incites others to commit genocide," drawing directly on the precedent established in the ICTR's Media Case discussed above.[23] This provision requires proof of direct and public incitement specifically to genocide — a narrower category than hate speech or incitement to "mere" discrimination or hostility — and requires the underlying crime of genocide, or an attempt at it, to be genuinely at issue, meaning it applies only in the most extreme atrocity contexts rather than as a general-purpose hate speech prohibition.

International humanitarian law's Additional Protocols to the Geneva Conventions do not directly regulate hate speech as such, but customary IHL principles concerning the protection of civilians have been invoked by some commentators to argue that dehumanising propaganda which measurably increases the risk of attacks against protected civilian populations during armed conflict may itself constitute a violation of the broader protective object and purpose of the Conventions, though this remains a contested and underdeveloped area of the law rather than a settled doctrine.

E. Regional Human Rights Frameworks

Europe. The European Court of Human Rights has developed an extensive body of jurisprudence under Article 10 of the European Convention on Human Rights (freedom of expression), applying a proportionality analysis broadly consistent with the ICCPR framework, while also permitting member states a degree of latitude, termed the "margin of appreciation," in balancing expression against other Convention rights given the differing national contexts and histories of Council of Europe member states. The Court has, in a consistent line of jurisprudence including cases such as *Garaudy v. France* (2003), treated Holocaust denial and genocide denial as falling outside Article 10 protection entirely, applying the "abuse of rights" doctrine in Article 17 of the

Convention, which excludes from Convention protection any activity aimed at the destruction of the rights and freedoms set out in the Convention itself.[24] The European Union's Digital Services Act (2022) represents the most comprehensive binding platform-governance instrument currently in force anywhere in the world, imposing binding transparency, risk-assessment, and content-moderation obligations on designated "very large online platforms" (those with more than 45 million monthly users in the EU), including specific obligations relating to systemic risks arising from the amplification of illegal content and disinformation, particularly during elections — though its extraterritorial reach and suitability as a universal model remain genuinely contested among states outside the European regulatory tradition.[25]

The Inter-American System. The Inter-American Court of Human Rights, applying Article 13 of the American Convention on Human Rights, has developed jurisprudence generally regarded as among the most protective of expression of any regional system, reflecting the Latin American region's own history of state censorship under authoritarian governments; the Inter-American Commission's Office of the Special Rapporteur for Freedom of Expression has been particularly active in documenting and cautioning against the use of disinformation legislation to target journalists across the region.

The African System. The African Commission on Human and Peoples' Rights has adopted the Declaration of Principles on Freedom of Expression and Access to Information in Africa (2019), which explicitly addresses digital-era concerns including internet shutdowns (a documented and recurring practice by several African governments during periods of political unrest or elections) and platform content moderation, situating the African regional framework as one that must simultaneously address both the historical legacy of colonial-era press restriction and the specific harms documented in cases such as the Rohingya crisis, given the significant number of African states that, like Myanmar, experienced rapid, platform-dominated internet adoption without the intervening decades of independent print and broadcast media development seen in other regions.

F. Comparative Domestic Frameworks

Delegates should also be aware of significant divergent national approaches, since these often shape a delegation's negotiating position more directly than the international instruments above.

Germany's Network Enforcement Act (NetzDG, 2017) was among the first national laws to impose binding takedown-timeline obligations on large platforms (requiring removal of "manifestly unlawful" content within 24 hours), and has served as a template — cited both approvingly and critically — for subsequent legislation in other jurisdictions.

India's Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 impose extensive obligations on "significant social media intermediaries," including traceability requirements for encrypted messaging services, and have been the subject of significant domestic and international human rights criticism regarding their compatibility with Article 19 standards, alongside government defence of the rules as necessary and proportionate given documented harms including the WhatsApp-linked lynching crisis discussed elsewhere in this guide.

Singapore's Protection from Online Falsehoods and Manipulation Act (POFMA, 2019) empowers government ministers to issue correction directions or takedown orders in respect of statements they determine to be false and contrary to the public interest, an approach the Singaporean government defends as a targeted and proportionate tool subject to judicial review, and which critics, including international press freedom organisations, characterise as vesting excessive unilateral authority in the executive branch to determine truth.

The United States' Section 230 of the Communications Decency Act (1996) takes a markedly different regulatory philosophy, providing broad platform immunity from liability for user-generated content, reflecting a policy judgment that intermediary liability would incentivise excessive removal of lawful speech; this framework remains subject to ongoing domestic political and legislative debate regarding potential reform.

China's Cybersecurity Law and related regulations impose extensive content-control obligations tied to a state-centric conception of "network sovereignty," requiring platforms operating in China to remove content the state deems to threaten social stability or state security, reflecting a fundamentally different starting premise regarding the relationship between the state and permissible online expression than the rights-based frameworks discussed above.

This comparative divergence is not incidental background — it is the empirical foundation of the bloc positions discussed later in this guide, and delegates should draw on the specific national framework of their assigned state wherever relevant.

---

[19] International Covenant on Civil and Political Rights, 16 December 1966, Article 19. [20] Human Rights Committee, General Comment No. 34 on Article 19, UN Doc CCPR/C/GC/34 (2011). [21] OHCHR, *Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred*, A/HRC/22/17/Add.4 (2012). [22] International Convention on the Elimination of All Forms of Racial Discrimination, 21 December 1965, Article 4. [23] Rome Statute of the International Criminal Court, 17 July 1998, Article 25(3)(e). [24] European Convention on Human Rights, Article 10; *Garaudy v. France*, ECtHR App. No. 65831/01 (2003). [25] European Parliament and Council, Regulation (EU) 2022/2065 (Digital Services Act), OJ L 277/1.

GAPS IN THE EXISTING REGULATORY FRAMEWORK

The gaps in the existing regulatory framework are not peripheral or marginal; they are structural and systemic. They collectively define the governance challenge that this committee is tasked with addressing.

A. The Jurisdictional Vacuum Over Private Platforms

The most foundational gap is jurisdictional. International human rights law, as traditionally construed, binds states, not private corporations. Article 19 and Article 20 of the ICCPR impose obligations on states parties to protect and restrict expression respectively, but the entities that today make the overwhelming majority of practical decisions about what speech is amplified, suppressed, or removed are private platform companies operating global services from a small number of jurisdictions, largely under their own privately drafted terms of service rather than any binding international human rights standard. The UN Guiding Principles on Business and Human Rights (2011) articulate a corporate "responsibility to respect" human rights, including a due diligence obligation, but this responsibility is explicitly framed as distinct from, and weaker in enforceability than, the binding legal obligations imposed on states, and lacks any direct enforcement mechanism at the international level.[26] This gap means that the entities with the greatest practical power over the modern information environment are subject to comparatively limited binding international legal constraint, though this is beginning to change through instruments such as the EU Digital Services Act, discussed above.

B. The Absence of a Binding Disinformation Framework

Unlike hate speech, which is at least partially addressed by Article 20(2) ICCPR and Article 4 ICERD, disinformation has no dedicated binding international instrument at all. States seeking to combat disinformation must rely on adjacent legal frameworks — electoral law, consumer protection law, defamation law — none of which was designed for the problem and each of which captures only a fragment of it. This absence has produced two divergent and, according to many observers on different sides of the debate, equally troubling state responses: in some jurisdictions, largely voluntary and inconsistent platform self-regulation with minimal binding transparency requirements; and in a growing number of other jurisdictions, sweeping "fake news" or "false information" criminal laws that the UN Special Rapporteur on freedom of expression and multiple press freedom organisations have documented as being disproportionately used to prosecute journalists, opposition politicians, and government critics rather than genuine disinformation actors.[27] Delegates should note that governments defending such laws generally dispute this characterisation, arguing that documented misapplication in specific cases does not invalidate the underlying regulatory objective, and that judicial oversight mechanisms within these laws provide adequate safeguards — a position civil society critics contest.

C. The Encryption and Private Messaging Gap

A substantial and growing proportion of disinformation and hate speech dissemination now occurs through end-to-end encrypted messaging applications such as WhatsApp, Signal, and Telegram, which by design are inaccessible to the platform-level content moderation, algorithmic transparency, and fact-checking mechanisms developed for public social media. India's 2018 WhatsApp-driven mob violence crisis, in which viral encrypted forwards falsely claiming child-kidnapping gangs were operating in specific villages contributed to at least twenty documented lynchings within a matter of months, illustrates the acute and potentially lethal character of this gap.[28] No existing framework offers a widely accepted, rights-respecting mechanism for addressing harmful content within end-to-end encrypted environments, because any technical mechanism capable of scanning encrypted content for prohibited material necessarily creates a vulnerability that can be repurposed for other forms of surveillance, generating a direct and, to date, unresolved conflict between the right to privacy (protected under Article 17 ICCPR) and the imperative to prevent incitement and harmful disinformation. WhatsApp's own response — imposing strict limits on the number of times a message can be forwarded — represents one design-level intervention that does not require breaking encryption, though its effectiveness at scale remains debated among researchers.

D. Attribution and the Anonymity Problem

Attribution of coordinated disinformation campaigns to specific state or non-state sponsors is technically and legally fraught. Platforms can identify coordinated inauthentic behaviour with reasonable confidence through network analysis, but formally and publicly attributing such campaigns to a specific government — with the diplomatic and legal consequences that follow — requires an evidentiary standard that is rarely met with full confidence outside the clearest cases, as the extensive but frequently contested attribution debates surrounding numerous alleged state-linked influence operations illustrate.[29] This attribution gap has a direct legal consequence: it creates significant practical accountability limitations for state-sponsored disinformation operators, since the mechanisms of state responsibility under international law — including those codified in the International Law Commission's Articles on Responsibility of States for Internationally Wrongful Acts (2001) — generally cannot be triggered against an unattributed or only tentatively attributed actor, and the "effective control" standard for attributing the conduct of proxy or affiliated non-state actors to a state, articulated by the International Court of Justice in the *Nicaragua* case, sets a demanding evidentiary bar that many documented influence operations do not clearly meet.[30]

E. The AI Governance Gap

There is currently no multilateral, universally binding instrument governing the development, deployment, or labelling of generative AI systems capable of producing synthetic disinformation or hate content at scale. The Council of Europe's Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, opened for signature in 2024, represents the most significant binding step taken to date, but its membership is regional rather than universal, and its provisions relating specifically to synthetic media are framed at a relatively high level of generality, leaving substantial implementation discretion to individual states parties.[31] The absence of binding, universal provenance and labelling standards for AI-generated content means that the technology best positioned to industrialise disinformation production at global scale is currently among the least internationally regulated components of the entire ecosystem.

F. The Linguistic and Resourcing Inequity Gap

As documented extensively in the Rohingya case study, content moderation resources — both automated classifier training data and human review capacity — are heavily concentrated in English and a small number of other high-resource languages. Independent research and platform-disclosed data have repeatedly found that automated hate speech detection systems perform measurably less accurately in lower-resource languages, and that platforms employ disproportionately fewer human moderators per user for these languages, creating a documented global inequity in which populations in linguistically and economically lower-resourced states or regions face substantially higher practical exposure to unmoderated hate speech and disinformation than users in wealthier, higher-resource-language markets, despite formally identical platform policies.

G. The Cross-Border Jurisdiction and Enforcement Gap

Content that is unlawful in one jurisdiction is frequently lawful, or hosted by a platform with no meaningful physical or legal presence, in another, creating significant practical enforcement difficulties for any single state seeking to act against extraterritorial disinformation or hate speech sources. The absence of any binding multilateral mutual-legal-assistance framework specifically tailored to this agenda means that states are frequently limited to blunt instruments — such as blocking access to specific platforms or content within their own territory, an approach that itself raises significant Article 19 concerns given its breadth — rather than more targeted forms of cross-border cooperation.

---

[26] UN Human Rights Council, *Guiding Principles on Business and Human Rights*, A/HRC/17/31 (2011). [27] UN Human Rights Council, Report of the Special Rapporteur on freedom of opinion and expression, A/HRC/47/25 (2021). [28] BBC News investigations into WhatsApp-linked mob violence in India (2018); Banaji & Bhat, *WhatsApp Vigilantes* (LSE, 2019). [29] US Department of Justice, "Mueller Report" (2019), Vol. I; see also

ongoing, contested attribution debates regarding other alleged state-linked operations documented in Meta's Adversarial Threat Reports. [30] International Law Commission, Articles on Responsibility of States for Internationally Wrongful Acts, UN Doc A/56/10 (2001), Article 8; *Military and Paramilitary Activities in and against Nicaragua* (Nicaragua v USA) [1986] ICJ Rep 14, para 115. [31] Council of Europe, Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, CETS No. 225 (2024).

BLOC POSITIONS AND GEOPOLITICAL DYNAMICS

The following overview presents the general policy tendencies documented in the public positions, national legislation, and multilateral negotiating record of the states and groupings listed. These are broad characterisations rather than fixed positions, and delegates representing specific states should conduct additional country-specific research, since individual states within each grouping frequently diverge from the general tendency described.

A. Western States (United States, European Union member states, United Kingdom, Canada, Australia)

Western states are generally strong rhetorical defenders of Article 19 protections and are institutionally cautious regarding binding international content-prohibition instruments, reflecting constitutional traditions — particularly the First Amendment tradition in the United States — that place a comparatively high threshold on permissible restriction of expression. The European Union, by contrast, has moved furthest of any grouping toward binding platform-governance regulation through the Digital Services Act and the voluntary-turned-binding Code of Practice on Disinformation, reflecting a European legal tradition generally more comfortable with proportionate restriction of expression in service of dignity, non-discrimination, and democratic integrity, and directly informed by the region's twentieth-century history of propaganda-fuelled atrocity. This creates a notable transatlantic divergence within the broader Western grouping: EU member state delegates are more likely to advocate for binding platform obligations and defined content-removal timelines, while delegations aligned with the US constitutional tradition are more likely to favour voluntary industry codes, media literacy investment, and narrower, more precisely defined restriction categories.

B. Russian Federation and Aligned States

Russia has been identified in multiple independent investigations, national intelligence assessments, and platform-disclosed takedown data as a documented source of large-scale coordinated disinformation operations, a characterisation Russia itself has consistently and publicly disputed. Simultaneously, Russia has advocated at the UN General Assembly and the Human Rights Council, including through its periodic sponsorship of resolutions on "information security," for expansive state-sovereignty-based frameworks that would grant states broad domestic authority to restrict online content characterised as false or destabilising to social order. Western states and a number of civil society organisations have characterised this advocacy as a vehicle for legitimising domestic content control rather than a genuine disinformation-countering measure; Russia and aligned states dispute this characterisation and frame their position as a defence of legitimate state sovereignty over domestic information space against what they characterise as externally imposed liberal-Western normative frameworks. In SOCHUM, this grouping is likely to support strong state-sovereignty language and to resist binding platform-transparency obligations framed around Western regulatory models.

C. People's Republic of China

China maintains one of the most extensive domestic internet content-control apparatuses of any state and has consistently framed content regulation as a matter of national sovereignty and social stability rather than a question subject primarily to international human rights scrutiny. China has advocated at the UN, including through the International Telecommunication Union and other technical bodies, for a "cyber sovereignty" model of internet governance that would grant states primary and largely unreviewable authority over content within their own jurisdiction, generally resisting international human-rights-based frameworks that it and aligned states characterise as vehicles for external political interference in domestic governance. Chinese delegates in SOCHUM are likely to support language emphasising state authority over platform content, capacity-building and technology-transfer provisions, and to resist resolution language framed primarily around Article 19 protections, independent civil-society oversight mechanisms, or platform obligations that would apply extraterritorially to domestic Chinese platforms.

D. Non-Aligned Movement and the Global South

Developing states have historically approached this agenda through a dual lens of protection and capacity, and this grouping is characterised by more internal diversity than any other discussed here. Many states within this grouping have experienced acute, sometimes lethal, documented harms from hate speech and disinformation — the Rohingya crisis in Myanmar and the WhatsApp-linked violence in India being the most extensively documented examples — and their delegates are frequently strong advocates for meaningful platform accountability and resourcing equity. Simultaneously, a significant number of states within this same grouping have enacted "fake news" or similar legislation that international press freedom organisations have documented as being used, in specific instances, against journalists and political opponents, and delegates from this grouping are frequently wary of international frameworks that could be invoked, whether by their own or other governments, to justify further domestic censorship, or that impose compliance costs disproportionately borne by smaller platforms, lower-resourced governments, and civil society actors with limited legal capacity. States within this grouping are also likely to raise the linguistic and resourcing inequity of current content moderation as a matter of global digital equity, and to press for capacity-building, technology transfer, and equitable resource allocation as a condition of accepting new international obligations.

E. Small Island Developing States and States with Limited Digital Infrastructure

A distinct sub-grouping, including many Pacific and Caribbean small island states, faces this agenda from a position shaped less by domestic platform-content control ambitions than by acute vulnerability — limited domestic content moderation capacity, near-total dependence on foreign-owned platforms with no meaningful local presence or accountability, and, in several documented instances, disinformation campaigns linked to broader geopolitical competition for regional influence. These states are frequently strong advocates for binding platform obligations precisely because they possess the least unilateral leverage to compel platform compliance through domestic regulation alone, and are among the strongest supporters of multilateral, UN-level solutions to this agenda as opposed to purely national or regional regulatory approaches.

F. States with Documented Recent Atrocity or Mass-Violence Experience

States including Rwanda, and states within the wider Great Lakes and Southeast Asian regions with direct historical experience of media-fuelled mass violence, bring a distinct moral and evidentiary authority to this agenda, frequently citing their own national experience — and the ICTR's Media Case jurisprudence specifically — in support of a more expansive reading of the Article 20(2) incitement threshold and stronger platform accountability obligations, while also, in some cases, facing their own domestic criticism regarding the compatibility of post-genocide-era media regulation with contemporary Article 19 standards.

PREVIOUS UNITED NATIONS ACTION

The UN General Assembly has repeatedly reaffirmed Article 19 and Article 20 of the ICCPR through periodic resolutions on freedom of expression and the safety of journalists, but has not to date produced any binding instrument specifically addressing hate speech and disinformation as a combined phenomenon. The UN Strategy and Plan of

Action on Hate Speech, launched by the Secretary-General in 2019, established a system-wide framework for the UN's own institutional response, built around thirteen commitments spanning monitoring, education, and partnership with civil society and the private sector, but is explicitly framed as non-binding policy guidance rather than a legal instrument; its accompanying 2020 Guidance Note on addressing hate speech through education and other measures has similarly not been codified into treaty law.[32]

The Human Rights Council's establishment of the mandate of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has produced a substantial and influential body of thematic reporting, including dedicated reports on online hate speech (2019), disinformation and freedom of opinion and expression in the context of electoral processes (2021), and the regulation of user-generated online content (2018), which collectively constitute the most authoritative soft-law elaboration currently available of how existing ICCPR standards should apply to digital platforms.[33] The Special Rapporteur's reports have consistently emphasised a "human rights by design" approach requiring platforms to conduct human rights due diligence consistent with the UN Guiding Principles on Business and Human Rights, and have repeatedly cautioned against content-removal mandates that, in effect, outsource core judicial functions — determining what speech is unlawful — to private companies without meaningful judicial oversight, transparency, or avenue for appeal.

UNESCO's Guidelines for the Governance of Digital Platforms (2023) represent the most detailed multilateral elaboration to date of proposed platform obligations relating to disinformation, proposing risk-based regulatory categories differentiated by platform size and reach, mandatory transparency reporting on content moderation decisions and algorithmic systems, and structural support for independent fact-checking organisations, though these guidelines remain voluntary and have not been adopted as binding treaty law by the General Assembly or any other UN political organ.[34]

The Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online, launched in 2019 by the governments of New Zealand and France following the livestreamed Christchurch mosque shootings, represents a narrower but operationally significant multi-stakeholder commitment specifically targeted at the intersection of extremist violence and online amplification, bringing together a coalition of governments, technology companies, and civil society organisations, though — like the UNESCO guidelines — it remains a voluntary political commitment rather than binding international law, and notably several major states, including the United States (at the time of the Call's launch) and a number of other significant digital-policy actors, declined to join.[35]

The Security Council has addressed narrow, adjacent aspects of this agenda within its own mandate, most notably through Resolution 2354 (2017), which called upon member states to develop and implement comprehensive counter-narrative and communication strategies to counter terrorist propaganda, situating incitement to terrorism specifically — rather than hate speech or disinformation as general categories — within the Council's threat-to-international-peace-and-security framework.

---

[32] United Nations, *United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences* (2020). [33] UN Human Rights Council, Reports of the Special Rapporteur on freedom of opinion and expression, A/73/348 (2018); A/74/486 (2019); A/HRC/47/25 (2021). [34] UNESCO, *Guidelines for the Governance of Digital Platforms* (2023). [35] Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online (2019).

KEY ARGUMENTS AND ANALYTICAL TENSIONS

The following section presents the principal analytical tensions this committee must navigate. Each is presented with the strongest available arguments on multiple sides. Delegates should not treat any single position below as the "correct" or Executive-Board-endorsed answer; the purpose of this section is to equip delegates to construct rigorous arguments consistent with their assigned state's policy tradition, not to signal a preferred outcome.

A. Binding New Law Versus Better Enforcement of Existing Law

One school of thought holds that the existing ICCPR/ICERD framework is substantively adequate but poorly enforced, and that Article 20(2) ICCPR and Article 4 ICERD already provide sufficient legal basis to prohibit online incitement; on this view, what is missing is not law but state political will, platform compliance infrastructure, and international coordination on enforcement, meaning the committee's task should be primarily institutional — strengthening reporting mechanisms, building platform-transparency obligations, and operationalising the Rabat six-part threshold test — rather than legislative innovation.

A competing school of thought holds that the ICCPR's fundamentally state-centric obligations structure is poorly suited to a world in which private platforms, not states, control content visibility at the scale relevant to this agenda, and that meaningfully new binding instruments are required to bind platforms directly to human-rights-consistent content moderation standards, with enforceable mechanisms and penalties for non-compliance analogous to those found in instruments such as the EU Digital Services Act.

Both positions must confront a shared and serious risk: any new binding obligation to restrict content, however carefully drafted, provides a ready legal justification that some states may invoke to suppress legitimate dissent — a pattern already extensively documented in existing national "fake news" legislation. Delegates advancing either position should be prepared to explain what specific safeguards, if any, their proposed approach would build in to mitigate this risk.

B. The Encryption Dilemma

A sharp and largely unresolved tension exists between combating the real, documented, and in some cases lethal harms of encrypted disinformation (as in the WhatsApp-linked violence in India) and preserving the privacy and security protections that end-to-end encryption provides to journalists, human rights defenders, and ordinary users, particularly in contexts where state surveillance is itself a documented tool of repression. Proponents of stronger intervention argue that the demonstrated real-world harm justifies targeted technical measures, potentially including forms of content scanning, subject to strict safeguards. Opponents argue that any technical mechanism enabling detection of harmful content within encrypted channels necessarily creates a vulnerability that can be repurposed by state or non-state actors for mass surveillance, and that the resulting harm to privacy and associated rights (freedom of association, protection of sources) would outweigh the benefit, pointing instead to design-level interventions that do not require breaking encryption, such as WhatsApp's own forwarding limits introduced in response to the Indian crisis. Delegates should research their assigned state's specific position on encryption policy, since this often diverges from that state's general position on the broader hate speech and disinformation agenda.

C. Content-Level Regulation Versus Algorithmic Design Regulation

A structurally significant question for this committee is whether governance should focus on individual pieces of content (the traditional legal object of both incitement law and defamation law) or on the algorithmic systems that determine which content is amplified, regardless of whether any individual post is independently unlawful. Advocates of algorithmic-design regulation argue that because engagement-optimised recommendation systems have been documented to systematically privilege emotionally activating hate speech and disinformation over neutral content, content-level regulation alone cannot address the harm, and that platforms should be required to conduct and disclose algorithmic risk assessments, an approach embedded in the EU Digital Services Act's "systemic risk" provisions.

Critics of this approach — including some who otherwise support robust content moderation — note that the empirical evidence regarding algorithmic amplification's real-world causal effect on polarisation and belief in false claims is more contested than commonly assumed, citing large-scale 2023 randomised studies published in *Science* and *Nature* that found smaller effects than the "filter bubble" hypothesis predicts. They also argue that regulating algorithmic design necessarily requires regulators to make judgments about permissible engagement-optimisation strategies generally, a far more intrusive and commercially significant intervention than targeted content removal, and one that risks entangling regulators in essentially unreviewable technical and commercial decisions about platform architecture, extending well beyond the traditional competence of human rights law.

D. Freedom of Expression as Shield or Sword: The Weaponisation Concern

Every regulatory tool proposed in this agenda carries a documented dual-use character. The same six-part Rabat threshold test that can be used to identify genuine incitement to atrocity can, if adopted into domestic law without independent judicial oversight, be applied by a government to criminalise criticism of its own conduct by characterising such criticism as "hostility." The same platform take-down obligations that can remove genuinely dangerous synthetic disinformation can be invoked by a government to compel removal of documentary evidence of human rights abuses on the basis that such evidence is "fake" or "manipulated." This is not a peripheral or hypothetical concern; the UN Special Rapporteur on freedom of expression has documented numerous instances of ostensibly anti-disinformation laws being applied against journalists and opposition figures rather than genuine bad-faith disinformation actors. States that have enacted such laws generally dispute this characterisation in specific instances, or argue that documented misapplication does not by itself invalidate the underlying regulatory model, provided adequate judicial safeguards exist. This tension — between the real, documented harms of unregulated hate speech and disinformation, and the equally real, documented harms of regulatory overreach — runs through nearly every specific proposal this committee is likely to consider, and delegates should expect it to be the central axis of debate.

E. Universal Standards Versus Contextual and Cultural Sovereignty

A further recurring tension concerns whether hate speech and disinformation standards should be genuinely universal, or whether meaningful deference should be given to states' own assessment of their social, historical, and political context in determining acceptable thresholds — a debate that echoes, but is analytically distinct from, longstanding broader debates in international human rights law regarding universalism and cultural relativism. Proponents of strong universal standards argue that the core harms at issue (incitement to atrocity, mass-casualty disinformation) are sufficiently grave and cross-culturally recognised to justify a genuinely universal minimum threshold, akin to the Rabat Plan's six-factor test. Proponents of greater contextual deference argue that the same words or images can carry radically different social meaning and risk profiles depending on a society's specific history, demographic composition, and current political conditions, and that a rigid universal standard risks both under-protecting some societies and over-restricting others.

F. Individual State Responsibility Versus Platform Responsibility

Finally, delegates must consider where primary responsibility should lie for addressing hate speech and disinformation: with the states in whose territory the harm occurs and whose citizens are affected, with the (often foreign-headquartered) platforms whose design and moderation choices shape the information environment, with the individual disseminators of content, or with some combination calibrated by role and capacity. Each allocation of responsibility carries distinct practical and normative implications, and existing frameworks — the state-centric ICCPR/ICERD architecture on one hand, the platform-centric EU Digital Services Act model on the other — reflect genuinely different underlying theories of where the point of most effective and legitimate intervention lies.

GLOSSARY OF KEY TERMS

Astroturfing — The practice of masking the sponsors of a message or organisation to make it appear as though it originates from, and is supported by, grassroots participants, when it is in fact centrally coordinated.

Coordinated Inauthentic Behaviour (CIB) — A term used by major platforms to describe networks of accounts that work in concert to mislead users about who is behind a given communication or campaign.

Deepfake — Synthetic audio, image, or video content generated or manipulated using artificial intelligence techniques (typically generative adversarial networks or diffusion models) to convincingly depict a person saying or doing something they did not say or do.

Disinformation — False information deliberately created and disseminated with the intent to deceive and cause harm.

Malinformation — Genuine, accurate information disseminated with the intent to cause harm, such as leaked private data.

Misinformation — False information shared without the intent to deceive or cause harm, typically by a person who sincerely believes it to be true.

Provenance (content authenticity) — Verifiable information about the origin, authorship, and editing history of a piece of media, increasingly implemented through cryptographic standards such as C2PA.

Rabat Plan of Action — A 2012 OHCHR-facilitated framework proposing a six-factor contextual test for determining whether expression constitutes incitement under Article 20(2) ICCPR.

Systemic Risk — A regulatory concept, most prominently codified in the EU Digital Services Act, referring to platform-level risks (such as algorithmic amplification of illegal content) that arise from a platform's design and operation as a whole, rather than from any single piece of content.

RECOMMENDED READING AND SOURCES

Primary Legal Instruments

1. International Covenant on Civil and Political Rights (1966), Articles 19 and 20
2. International Convention on the Elimination of All Forms of Racial Discrimination (1965), Article 4

3. Rome Statute of the International Criminal Court (1998), Article 25(3)(e)
4. European Convention on Human Rights (1950), Articles 10 and 17
5. American Convention on Human Rights (1969), Article 13
6. UN Guiding Principles on Business and Human Rights (2011)
7. EU Digital Services Act, Regulation (EU) 2022/2065
8. Council of Europe Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (2024)
9. ILC Articles on Responsibility of States for Internationally Wrongful Acts (2001)

Key Reports and Soft-Law Instruments

1. OHCHR, Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred (2012)
2. UN Strategy and Plan of Action on Hate Speech (2019) and Detailed Guidance for Implementation (2020)
3. Human Rights Committee, General Comment No. 34 on Article 19 (2011)
4. CERD Committee, General Recommendation No. 35 (2013)
5. UN Human Rights Council, Reports of the Special Rapporteur on Freedom of Opinion and Expression (2018–2021)
6. UN Fact-Finding Mission on Myanmar Report, A/HRC/39/64 (2018)
7. BSR, Human Rights Impact Assessment: Facebook in Myanmar (2018)
8. UNESCO, Journalism, 'Fake News' and Disinformation Handbook (2018) and Guidelines for the Governance of Digital Platforms (2023)
9. Oxford Internet Institute, Industrialized Disinformation: Global Inventory of Organised Social Media Manipulation (2021)
10. Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online (2019)
11. World Health Organization, Infodemic Management: An Overview of Infodemiology (2020)
12. ICTR, Prosecutor v. Nahimana, Barayagwiza and Ngeze ("Media Case") (2003)
13. African Commission on Human and Peoples' Rights, Declaration of Principles on Freedom of Expression and Access to Information in Africa (2019)

Comparative National Legislation

1. Germany, Network Enforcement Act (NetzDG, 2017)
2. India, Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules (2021)

3. Singapore, Protection from Online Falsehoods and Manipulation Act (2019)
4. United States, Communications Decency Act, Section 230 (1996)

QARMA — QUESTIONS A RESOLUTION MUST ANSWER

Delegates drafting resolutions on this agenda should ensure their working papers and draft resolutions substantively engage with each of the following questions. These questions do not presuppose any particular answer; they are intended to test the rigour, coherence, and practical viability of any proposed framework.

1. How should hate speech and disinformation be defined and distinguished from protected expression, satire, and legitimate criticism?

2. Where should the line be drawn between restricting speech and protecting freedom of expression, and who decides?

3. Who bears primary responsibility for addressing this issue — states, platforms, individuals, or some combination — and how is that responsibility enforced?

4. How should encrypted and private communications be treated, given the tension between privacy and harm prevention?

5. Should algorithmic amplification itself be regulated, separately from individual pieces of content?

6. How should AI-generated synthetic media be addressed, given the limits of detection technology?

7. What safeguards can prevent any proposed framework from being used to suppress dissent, journalism, or minority voices?

8. How should the resolution address global inequities in content moderation capacity and resourcing?

9. Should the resolution work through existing UN mechanisms or propose new institutional architecture, and how does it relate to existing binding obligations under the ICCPR and ICERD?

10. How will implementation be monitored, and what accountability exists if a state under-enforces or overreaches?