GENEVA ACADEMY
Académie de droit international humanitaire et de droits humains
Academy of International Humanitarian Law and Human Rights

RESEARCH BRIEF

DIGITAL DISINFORMATION OPERATIONS: PART I – SYNTHETIC FORCES VS. HUMANS AND HUMAN RIGHTS

ABSTRACT

Digital disinformation operations (disinfo-ops) employ sophisticated synthetic activity to distort public discourse, manipulate democratic processes, and erode trust across societies. Unlike traditional forms of propaganda, these operations amplify false or misleading content at an unprecedented scale and precision, leveraging automated bots, troll networks, and AI-generated media to shape perceptions and suppress genuine voices. This report examines the mechanisms behind disinfo-ops and their impact on human rights, particularly freedom of expression, opinion formation, and the right to seek, receive and impart information. Removing inauthentic activity is a critical step to preserving human rights and protecting human agency in an increasingly polluted information ecosystem.

DECEMBER 2024 | STEVEN J. BARELA

This publication has been externally peer-reviewed

THE GENEVA ACADEMY A JOINT CENTER OF
GENEVA GRADUATE INSTITUTE
INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT
GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES
UNIVERSITÉ DE GENÈVE
FACULTY OF LAW

INTRODUCTION

The online information space has always evolved haphazardly, shaped by competing interests of individuals, organizations and commercial entities. Today it is profoundly disordered. Technical capacities allow for content – correct or incorrect – to be generated quickly; it takes much more time to produce something that closely aligns to the facts than it does to disseminate falsehoods. Marshaling experts to fact-check vast amounts of information online is already difficult, and there is frequent debate over the best way to describe reality, even among specialists. This is a normal part of both the democratic and scientific process. As a result, confusion persists among the general public about the value of the many different types of virtual information to be found. Efforts to teach people how to navigate the credibility of information in cyberspace through digital literacy programs are only beginning and come with their own paradoxes.[1] Beyond this, the algorithms that curate our newsfeeds are not built for accuracy, but rather for grabbing attention and holding engagement. All of this makes for a volatile information ecosystem that does not favor reliability and truth. Some have labelled it a "crisis of information."[2]

Compounding this disorder are nefarious actors who wish to sow confusion across borders within this fertile ground. Such intentional action is not new. The idea of weakening an adversary by planting falsehoods in local sources has a deep history in the 20th century.[3] Wide participation on social media platforms with an ease of content delivery has allowed for the rapid amplification of purposefully false narratives. The digital era has transformed the way societies interact with information, and online platforms provide tools to shape our understanding of the public sphere like never before. The combined effects of mis- and disinformation have far-reaching societal impacts, making it essential to map out the mechanics of this disruptive force. Crucially, much of this digital activity is not organic but artificially generated – driven by bots, inauthentic actors and algorithms – highlighting a structural vulnerability that can be addressed through policy and regulatory measures.

From a legal perspective, digital disinformation presents significant challenges to the international community.[4] To begin to chart the complex applicable international law, it is important to highlight the now widely accepted interpretation that the principles that flow from sovereignty and the UN Charter indeed apply in cyberspace.[5] Different areas of international law, including the protection of fundamental human rights, the rules governing conflict, and principles shaping State relations and self-determination, are implicated here. There are also the critical questions of responsibility and accountability, particularly in contexts where anonymity and plausible deniability complicate the attribution of harmful actions. Together, these elements provide a foundation for understanding how international legal obligations intersect with the evolving difficulty of digital disinformation.

To enhance accessibility, this research has been divided into four interrelated reports, each addressing a distinct aspect of the danger posed – referred to in this report series as "disinfo-ops". Part I, presented here, examines the fundamental mechanisms involved behind these operations and challenges the idea that freedom of expression is in direct conflict with efforts to mitigate harm. Instead, it highlights how synthetic interference is often deliberately engineered to distort discourse, suppress genuine voices, and manipulate opinion formation. Part II provides a detailed examination of the other legal dimensions under international law, including considerations of sovereignty, non-intervention, self-determination, and accountability. Additionally, two specialized Info-Briefs address critical thematic areas: one examines how the law of armed conflict treats deception, while the other explores the European Union's regulatory landscape in facilitating independent research through privacy-preserving data access. Together, these reports offer a comprehensive framework for understanding and addressing the evolving problem.

Setting the stage for the subsequent reports in this series, this report examines how online disinformation today exposes digital vulnerabilities, amplifies polarization and warps public discourse. While distorted material has long been a tool of manipulation, the degree and speed of its spread in the digital age pose new complications, particularly for protecting freedom of expression and opinion. This report unpacks how artificial activity – or synthetic forces – exploit algorithmic design and personal data for micro-targeted amplification, intensifying the issue and making effective responses difficult to formulate. By mapping these dynamics, we will clearly see why leading scholars have noted:

Although there is nothing necessarily new about propaganda, the affordances of social networking technologies – algorithms, automation, and big data – change the scale, scope, and precision of how information is transmitted in the digital age.[6]

Addressing this complex challenge requires a strategic and adaptive approach akin to managing an ecosystem. Mis- and disinformation act as pollutants, contaminating public discourse and undermining trust across society.[7] Understanding the scope and impact of this degradation demands access to the data driving these mechanisms, enabling researchers to identify the precise components so that policymakers can devise effective interventions. Moreover, filtering out inorganic contamination is an essential first step to restoring the trust and integrity that underpin a thriving, human-centered information ecosystem.

MAPPING THE MIS- AND DISINFORMATION LANDSCAPE

UNPACKING THE CONCEPT

There is no international consensus on a definition for the term "disinformation."[8] Without precise terminology, addressing the harm caused by such an action becomes difficult and leaves room for manipulation and exploitation of legal grey areas.[9] By labeling any disliked material as disinformation or propaganda, the information landscape becomes littered with competing accusations causing people to lose faith in their leaders, along with their own capacity to navigate the morass.

Academics have created a taxonomy to distinguish between mis- and disinformation. Fundamentally, they can be understood as distinct in their intent – misinformation is false or misleading information spread without malice, and disinformation is deliberately deceptive.[10] Disinformation aims to manipulate public opinion to foster uncertainty and confusion. Chaos creates opportunities for malfeasance, subversion or misconduct. In contrast, misinformation – though not generated with harmful intent – often proliferates through unwitting individuals who share false information without realizing its inaccuracy or simply find it amusing. In other words, once deliberately manipulated content reaches the public, well-meaning parties may inadvertently share it, turning disinformation into misinformation as it spreads beyond its original, deceptive intent. This "snowball effect" boosts the content's visibility with each share, making it seem increasingly credible. Together, this blend of mis- and disinformation creates a thick fog that is extremely difficult to pierce in order to identify responsibility.[11]

Determining truth is inherently complex and ever evolving. Experts frequently disagree on the interpretations of facts and priorities. While this diversity of perspectives is essential for democratic and scientific progress, it also creates tension in managing an increasingly contested information space.[12] The term "disinformation" originates from the Russian dezinformatsiya, a concept developed during the Soviet era as part of state-sponsored campaigns designed to manipulate perceptions through a strategic blend of truth and falsehoods.[13] This tactic exploits the inherent difficulty of fully grasping objective reality, using uncertainty as a tool to construct misleading narratives. Meanwhile, the United States pioneered large-scale psychological operations (PSYOPs) during World War II, employing strategic influence campaigns to shape behavior and achieve political or military objectives.[14] Both laid the groundwork for modern

cognitive manipulation, but today's digital tools fueled by personal data have largely solved previous challenges in targeting and credibility, enabling influence campaigns to operate with unprecedented accuracy and impact.

Today, a variety of other terms are used to describe the phenomenon of false or misleading information. Some examples include "fake news," "manipulated content," "cognitive warfare," "influence operations," "propaganda" and "xuanchuan."[15] These terms often carry distinct meanings and can vary across disciplines and contexts. This report will not directly address this complexity; instead it will use the term "disinfo-ops" as a broad framework of disinformation operations to capture the wide scope of a knotted problem.[16] This choice emphasizes how the available mechanisms operate in a disordered information system. It should not, however, be understood as a conclusion on what is the best terminology to capture the evolving use of information and communication technologies (ICTs).

The borderless nature of the Internet has also given actors interested in producing and moving distorted communications an unprecedented reach into foreign communities. Individuals from one country can pretend to be local residents and engage in targeted political discourse to trigger reactions or push conversations in extreme or polarizing directions.[17] This can compound the impact of domestic groups that equally benefit from creating chaos and the upheaval caused by lies.[18] While selected governments and international organizations have responded to this threat by raising alarms and imposing regulations,[19] the cross-border nature of digital communication complicates enforcement efforts – especially when the desire to fuel discord and doubt is borne at home and supported abroad.[20]

MECHANISMS OF PROLIFERATION

Lies and falsehoods have always existed. Manipulation often follows. What is different today are the tools used to move untruths at speed and to inundate individuals with specific fabrications that are tailored to their interest or political preferences. The forces driving the proliferation of mis- and disinformation are made up of a complex interplay of both human actors and technological tools. Together, they accelerate and amplify the spread of false or misleading content – and obscure its origins. Key among these devices is the collection of personal data, automated bots, malicious trolls/State-sponsored cyber-armies, deepfakes, and corporate algorithms. Each plays a distinct role in shaping the online information space and collectively magnify the impact of harmful narratives. Moreover, it is necessary to understand that these instruments are being used across various online platforms, making cross-platform research essential.[21] Yet, without data access for scientific study, we remain unable to grasp the full extent of how these tools shape the online environment or the sophistication with which they are deployed.[22]

It is also important to note that much of this taint spreads through inauthentic activity sponsored by States or monied interests.[23] Much like environmental pollution disrupts natural systems, this tech-driven surge of synthetic activity undermines genuine human interaction. It drowns out and overwhelms organic discourse. In contrast, authentic expression, even when it challenges or creates discomfort, is a fundamental human right that propels societal progress. Strictly limiting counterfeit voices, magnified by automation, arguably preserves the integrity of earnest exchange between real people.[24]

What follows is a focused presentation of the principal mechanisms.

- **Collection of Private Data: fuel for targeted manipulation.** The quantity of digital data being generated daily is immense.[25] Large amounts of it can be tied to individuals, the collection of which began for the purpose of trained marketing campaigns – otherwise known as micro-targeting.[26] This capacity has been repurposed for highly precise political influence operations.[27] Platforms and third-party actors collect vast amounts of private data, including browsing habits, location, interests, and demographic details, which can then be analyzed to create detailed user profiles.[28] This personal data summary enables tailored messaging that taps into individual biases, fears or political bents to increase the likelihood of engagement – thus the spread of specific content to receptive audiences. By feeding into the mechanics described below, personal data collection sharpens the precision and impact of particularized disinfo-ops, heightening the risk that it is unknowingly spread as misinformation.[29] Most concerning, this has largely resolved one of the key challenges for effective PSYOPs and dezinformatsiya campaigns – understanding a target and orienting credible content to manipulate it.

- **Bots: automated networks of social media accounts.** Bots operate at scale and are designed to mimic human behavior. Networks of bots can be programmed

to rapidly deploy at critical moments to amplify content, including false or misleading information, by generating large volumes of posts or interactions (likes, shares, comments). This manipulates algorithms to further boost selected content. Their speed and efficiency make them a powerful tool for spreading distorted information across platforms with minimal human intervention.[30]

- **Trolls: malicious content provocateurs.** Trolls are individuals or groups who deliberately provoke or mislead others online, often by posting inflammatory or false content. This activity can manipulate public opinion, sow discord, and disrupt discussions by steering conversations toward divisive topics. They can engage in harassment or influence campaigns, contributing to the broader spread of harmful content – their impact is connected to the ability to assemble, direct and coordinate activity.[31] A hybrid phenomenon, often called "cyborgs," combines the efficiency of bots with the adaptability of human trolls to better avoid detection for more sophisticated manipulation.[32]

- **Troll Farms / Cyber-armies: coordinated digital forces.** These consist of organized groups, sometimes backed or indirectly supported by State actors, that engage in spreading disinformation for geopolitical or ideological purposes. Such digital forces can systematically manipulate online narratives, create confusion, and undermine trust in governmental processes or institutions – either domestic or international. Their activities range from social media manipulation to hacking, data leaks, and coordinated harassment campaigns. Attribution is notoriously difficult,[33] as the involvement of State actors is often obscured or denied, allowing them to distance themselves from the actions of such groups. This ambiguity not only complicates accountability but also shields States from direct repercussions.[34]

- **Deepfakes: manipulated visual deceptions.** AI-generated videos and images are becoming increasingly sophisticated, making it difficult to distinguish fabrication from reality. Deepfakes can convincingly depict individuals saying or doing things that did not take place, enabling the spread of false information, reputational harm, and political destabilization. While watermarking and detection tools provide some safeguards, the rapid evolution of deepfake technology has created an arms race where defensive measures struggle to keep pace, leaving individuals and smaller institutions unprotected.[35]

- **Algorithms: hidden engines of content amplification**. These engagement-driven curators are automated systems widely used across digital platforms, including social media, search engines, and streaming services, to prioritize content based on user interaction. Closely connected are 'recommender systems' that propose additional content to users.[36] Designed to boost participation, they often favor sensational or emotionally charged material, accelerating the spread of mis- and disinformation. As like-minded people form groups online, the reverberation of shared content has given rise to what have been called "filter bubbles and echo chambers."[37] Since the algorithms and recommender systems are closely guarded trade secrets, a complete understanding of their inner workings remains out of reach.[38]

This tangled web of humans and technology not only distorts public discourse but also obstructs efforts to address the problem at its root. Moreover, the development of generative AI technology means that these mechanisms of hidden persuasion have the potential to be applied at an even wider scale. In the words of one prominent expert: "people produce and consume lies, but the algorithms, data sets, and information infrastructure determine the impact of those lies."[39]

IDENTIFYING THE TARGETS: WHO SUFFERS?

When falsehoods permeate society, they undermine confidence and stability at every level. Nonetheless, particular groups and fields of practice are more susceptible than others. Marginalized communities are particularly vulnerable and often victims of harmful content exacerbating existing inequalities. Health and science can face serious consequences, as a disarray of information erodes public trust in evidence-based guidance. Journalists and media outlets tasked with providing reliable information can be discredited or attacked, further weakening public discourse. Moreover, while every society needs trustworthy sources of information, its absence is a particular detriment to the democratic processes that rely on authentic contestation, discussion and deliberation to

decide who will hold authority within a constitutional order. In short, mis- and disinformation contribute to an erosion of trust – in institutions, governments and even within interpersonal relationships – corroding social cohesion and deepening divisions within societies.

Marginalized Communities and Gendered Disinformation

Communities around the globe face systemic barriers to full inclusion in society, leaving them vulnerable to exclusion, discrimination, and targeted manipulation. Marginalized groups are often disproportionately impacted by digital disinfo-ops, as they become easy targets for narratives that exploit existing social, ethnic, or religious tensions. False or misleading content can be used to incite division and promote stereotypes. In extreme cases, such campaigns can even contribute to acts of violence. Limited access to reliable information often exacerbates these challenges, making it harder for marginalized groups to fully participate in civic and social life.[40]

A particularly insidious practice is gendered disinformation, which disproportionately targets women, LGBTQ+ individuals, and other gender identity groups.[41] As noted by the UN Special Rapporteur on Freedom of Opinion and Expression, female politicians, journalists, and human rights defenders frequently face digital campaigns designed to silence and delegitimize them.[42] These attacks reinforce patriarchal norms, deter participation in public discourse, and exacerbate digital inequalities. By weaponizing information in this way, such campaigns erode fundamental rights and undermine efforts toward equality.

The devastating impact of targeted disinformation was starkly evident in the 2017 Rohingya crisis in Myanmar, where false narratives played an important role in inciting violence against ethnic minorities.[43] In 2024, the Independent Investigative Mechanism for Myanmar (IIMM) released a report detailing the Myanmar military's information campaign against the Rohingya population.[44] During the 2017 clearance operations, a covert network of Facebook pages, secretly operated by the military, systematically spread hate speech and incited violence. This network, linked to the military through shared administrators and IP addresses, posted inflammatory content, with over 10,000 posts identified as hate speech, promoting extreme anti-Rohingya sentiments and justifying acts of aggression. This case highlights how disinfo-ops can weaponize existing prejudices to escalate persecution.[45]

Health and Science

Health and science suffer when the nature of scientific consensus – established through peer review and rigorous replication checks – is corrupted. The scientific process embraces debate and dissent to achieve consensus. Yet this process can be distorted by amplifying fringe voices, making it appear as though credible evidence is evenly divided when, in reality, a large majority of scientific views point in the same direction. This tactic can cloud public understanding, making it difficult to distinguish between consensus-driven guidance and isolated opinions. When such manipulation takes place within receptive communities it weakens trust in scientific findings, allowing misinformation to thrive under the guise of legitimate scientific debate.[46]

The global community witnessed this phenomenon when the COVID-19 pandemic quickly evolved into an "infodemic." When the virus was beginning to spread in February 2020,[47] the World Health Organization (WHO) warned that mis- and disinformation was spreading in parallel to the virus itself.[48] From false cures to conspiracy theories about the virus's origins, the public was inundated with misleading content that fueled confusion and distrust in official health advice. WHO officials recognized early on that this flood of harmful information posed a serious danger to public health by undermining evidence-based guidance and amplifying fear. This wave of mis- and disinformation later became a major obstacle during the vaccine rollout, as falsehoods surrounding vaccine safety and efficacy fueled widespread hesitancy and slowed global immunization efforts.[49]

Journalism & Media

The spread of distorted information erodes the credibility of quality journalism, undermining public trust in the media's role as an essential, if not entirely unbiased, source. While good journalism strives to uphold objectivity, biases can emerge from individual perspectives, editorial choices, or broader institutional leanings. Information campaigns exploit these natural vulnerabilities. They disguise false content by mimicking reputable outlets, creating fake personas that spread fabricated stories, or posing as independent news sources with seemingly genuine, underrepresented perspectives. This harm can extend beyond the public perception of media sources to endanger journalists. Indeed, individuals committed to ethical standards and fact-based reporting can be targeted in such campaigns insofar as their work threatens a distorted narrative.[50]

Once respected for its journalism, *France-Soir* became an example of how disinformation can erode media credibility. The outlet was a leading French newspaper with a peak circulation of 1.5 million in the 1950s and 1960s. After going bankrupt in 2012, it re-emerged as an online-only entity in 2016. Criticized for publishing false information, *France Soir* was downgraded in 2020 by NewsGuard.[51] In 2022, it lost its official press credentials from France's *Commission paritaire des publications et agences de presse* (CPPAP), and an administrative court in Paris upheld this decision in 2024.[52] The time taken and financial costs incurred to arrive at this result demonstrate the challenge to traditional journalism and its protection in the current environment.

Legitimacy & Democracy

Legitimacy is essential for any government to function. Belief in the validity of the process that establishes an authority promotes compliance with its decisions and orders, enables collective action and stimulates cohesion within society. Disinfo-ops can threaten this legitimacy by distorting public perception of the process, eroding trust in institutions and government officials, and instilling doubt around the authority in power. Ultimately, by weakening a government's ability to engender trust and promote compliance with rules and policy, this manipulation undermines the critical balance needed for effective governance.[53]

Elections in the United States and Brazil have illustrated how manipulated information can severely disrupt electoral processes and undermine state authority. In the U.S., disinformation during the 2016 U.S. election, particularly through social media manipulation by foreign actors,[54] intensified political polarization and began to cast doubt on procedural integrity. Moreover, false narratives about electoral fraud led segments of the public to question the veracity of the election outcome, culminating in the attack on the Capitol on 6 January 2021.[55] Similarly, in Brazil, misleading narratives following the presidential election spurred unrest, with protestors storming government buildings in Brasília where over 1500 people were arrested.[56]

Trust

Society rests on a foundation of trust. It enables individuals to engage confidently in public life by fostering a sense of security, mutual respect, and shared purpose. When people are confident that systems are fair and public information is reliable, they are more likely to participate in and cooperate toward common goals.[57] In the digital age, such trust extends to the information ecosystem where people need to rely on the authenticity and reliability of the content they consume. As technologies and how humans use them complicate this landscape, establishing and maintaining trust in digital spaces (certainty of origin, intent, and accuracy of information) will become increasingly challenging,[58] and mechanisms for preserving public trust imperative.[59] One challenge in this regard is that "trust" is not a quantifiable entity. Examples will always be approximate and subject to criticism. Nonetheless, it can be seen that information campaigns actively undermine trust by using vulnerabilities within the information ecosystem, blurring the lines between fact and opinion, or accurate and erroneous information. Anonymity interwoven with malicious bots and troll armies in online spaces complicates this further. It allows malign actors to obscure the roots of content, amplify or drown out particular voices, and hide behind genuine freedom of expression – a right reserved for humans. While anonymity is essential for privacy and free expression, addressing its role in disinformation requires balancing these protections with accountability measures to impede the erosion of trust.[60]

ENGINEERED DISTORTIONS AND THE CHALLENGE TO HUMAN RIGHTS

The human right to freedom of expression and opinion is a building block of democratic society, fostering informed decision-making, open deliberation and the exchange of diverse ideas. Inserting deliberately false or misleading information into the public sphere of a self-governing society exploits these principles, polluting and distorting the information landscape while undermining trust in public discourse. This tension lies at the heart of contemporary legal and ethical debates. On the one hand, preserving freedom of expression requires protecting even unpopular or controversial speech; on the other, unchecked disinformation at scale can polarize societies, erode democratic processes, and harm fundamental rights. However, it is essential to note that freedom of expression is rooted in the organic exchange of ideas between individuals, not a contest for visibility against synthetic forces deployed at scale. Rather than an inherent legal contradiction, this tension is often artificially created – sometimes deliberately – by inauthentic actors who manipulate digital ecosystems to suppress and drown out human voices. Recognizing this dynamic shifts the focus, from restricting speech to addressing the structural vulnerabilities that enable such distortions, offering a more precise and rights-aligned path forward.

It is also important to note the collection and use of personal data presents a fundamental strain on another intersecting human right – the right to privacy.[61] The vast amounts of data gathered through tracking technologies on digital platforms and social media interactions are not only monetized for advertising but also exploited for targeted political campaigns.[62] By leveraging cookies and trackers embedded in websites to trace user interactions, browsing habits and online purchases, these systems use advanced analytics to build detailed profiles and curate highly personalized information environments.[63] The lack of transparency and meaningful user consent in these processes raises serious concerns to a number of human rights and have been identified by the Office of the High Commissioner of Human Rights as "inextricably linked to the personal data that powers the engines of digitized societies".[64]

Human rights law as embodied in Article 19 of both the Universal Declaration of Human Rights (UDHR)[65] and the International Covenant on Civil and Political Rights (ICCPR)[66] guarantees the right to freedom of opinion and expression. While the right to freedom of opinion is non-derogable and can never be legally breached, freedom of expression is not absolute. Restrictions on speech can be imposed when it endangers national security, public order, health, or the rights of others.[67] In such cases, restrictions must meet the strict tests of legality, necessity and proportionality.[68] Furthermore, speech or expression cannot rise to the level of incitement to hatred, discrimination and violence – all of which are prohibited under international law.[69] Disinformation often straddles these lines.

THE NON-DEROGABLE RIGHT TO FREEDOM OF OPINION

Freedom of opinion, as distinct from freedom of expression, is absolute and inviolable – even during states of emergency.[70] This provision guarantees that individuals are entitled to form and hold opinions without interference, emphasizing a total protection of the internal cognitive process and serving as a linchpin of personal autonomy. Unlike the right to freedom of expression, which is subject to certain limitations under Article 19(3) of the ICCPR, the freedom to hold opinions is fully shielded from inappropriate external intrusion or influence. This includes protection against coercion, penalization, or manipulation of an individual's opinion formation process. The right underscores a critical component of human independence, ensuring that individuals retain the capacity to think and form opinions freely, which is fundamental to personal integrity and democratic participation.[71] Importantly, during the negotiation of the ICCPR, there was ongoing debate and confusion about whether "freedom of thought"[72] and "freedom of opinion" were one and the same;[73] ultimately, both remained in the treaty.[74]

Disinfo-ops threaten freedom of opinion by warping the informational landscape upon which many individuals rely to form their views. While singular or even multiple instances of producing false content would not infringe on freedom of opinion, information operations often operate within a closed and bespoke information bubble that can distort individuals' ability to form views based on diverse, factual and reliable information. It is by flooding a personalized sphere with manipulated narratives that such campaigns skew perceptions. When amplified through algorithm-driven platforms, repeated exposure to falsehoods can entrench cognitive biases and powerfully shape opinions. This can compromise the autonomy of individuals to hold opinions free from undue influence, highlighting the need for robust safeguards to protect this

fundamental, non-derogable right. Such environments can also manipulate public discourse, suppress dissenting voices, and polarize societies, indirectly undermining the broader principles of democratic governance.

Both States and social media companies should prioritize this serious issue,[75] including by enabling and supporting data access for independent study. Research can help to fully understand the mechanisms of these campaigns, allowing the development of evidence-based strategies to protect freedom of opinion. Empirical insights can help reveal how manipulated narratives are disseminated and amplified, providing a basis for targeted safeguards. Moreover, innovative proposals should be advanced to ensure that only humans are interacting online and influencing each other;[76] inauthentic technologies should not be allowed to skew opinion formation.

THE RIGHT TO SEEK, RECEIVE AND IMPART INFORMATION

The ICCPR guarantees individuals have the right to seek, receive, and impart information of all kinds. They can do so across borders and through any type of media. This right also includes access to information that is disturbing, shocking or offensive,[77] and the accuracy of an opinion or interpretation is not grounds for prohibition.[78] This means that a society built on open debate must accept discomfort and disagreement, as long as it does not include the promotion of intolerance, unfair treatment, or physical harm.

To these ends, States have both negative and positive obligations under the human rights framework. The negative obligation requires States to refrain from actions that unduly restrict freedom of expression or suppress information. This includes avoiding Internet shutdowns, arbitrary censorship, and the misuse of laws to target critics or dissenting voices. Internet shutdowns, often imposed during elections or protests, disrupt access to information and are frequently deemed disproportionate under international human rights law.[79] Arbitrary censorship, whether through blanket bans or unrefined and undefined removals of content, undermines freedom of expression by silencing dissenting views without justification.[80] Misuse of criminal laws against "false information" often relies on vague language, allowing authorities to suppress legitimate criticism and dissent, contrary to international standards.[81] Restrictions on freedom of expression, such as those aimed at curbing disinformation, are permissible only under exceptional circumstances and must align with the tripartite test of legality, necessity, and proportionality.[82] This framework ensures that limitations are narrowly tailored and do not arbitrarily infringe on fundamental freedoms.

Positive obligations, by contrast, require States to take proactive measures to protect and promote the right to information. This involves ensuring the availability of diverse and accurate information, supporting independent journalism, protecting journalists, and addressing structural inequalities that may render certain groups more vulnerable to disinformation. States should adopt rights-respecting approaches that prioritize transparency, accountability, and public education. Fostering digital literacy can create an environment that empowers individuals to make informed decisions and participate meaningfully in deliberative processes.[83] Similarly, fostering collaboration with civil society, independent media, and international organizations can strengthen efforts to counter campaigns to distort information while safeguarding freedom of expression. By aligning these efforts with international human rights standards, States can address the complex challenges created while upholding the integrity of the information ecosystem. The free flow of authentic information generated by real humans can be a powerful tool for fostering a healthy information environment.[84]

RESPONSIBILITY OF DIGITAL PLATFORMS

Social media companies and digital platforms, while not bound by the same human rights obligations as States, are expected to respect human rights in their operations under the Guiding Principles on Business and Human Rights (UNGPs).[85] UN Special Rapporteurs on the promotion and protection of the right to freedom of opinion and expression have pressed online technology companies to adhere to human rights standards in all of their business practices.[86] The UNGPs emphasize the responsibility of businesses to avoid infringing on human rights and to address adverse impacts caused by their activities. However, the business models of these platforms, which rely heavily on advertising revenue, have come under scrutiny for exacerbating the spread of disinformation.[87] Algorithms designed to maximize engagement often prioritize sensational or polarizing content, amplifying harmful narratives while neglecting the broader societal impact.[88] Additionally, practices such as data harvesting and microtargeting exploit user information to create personalized content streams, which can reinforce echo chambers and deepen polarization.[89] These practices raise significant concerns

about user privacy and the obligations of companies to mitigate the harms their platforms facilitate.[90]

Efforts to address the problem through content moderation have similarly faced criticism for inconsistency and a lack of transparency.[91] Companies employ automated filters and manual reviews to identify and remove harmful content, yet these systems frequently fail to account for nuance, leading to over-removal of legitimate speech and under-removal of harmful material. Moreover, platforms often provide limited avenues for users to appeal moderation decisions or receive clear explanations of enforcement actions. Transparency reports, while a step in the right direction, often lack essential details about the reach and impact of disinformation, the effectiveness of content moderation policies, and the reliability of artificial intelligence tools.[92] Without adhering to higher standards of accountability, companies risk perpetuating distrust while falling short of their responsibility to respect human rights.

Beyond these shortcomings, a more fundamental issue remains largely unaddressed – the rampant proliferation of synthetic activity on digital platforms. Large platforms have little incentive to curb bot activity or users with multiple accounts, as "good" bots – those undetectable to automated detection systems – drive engagement, inflate user metrics, and ultimately increase advertising revenue. Any technical capability to reliably distinguish between authentic and synthetic users would challenge existing engagement-driven business models. Moreover, it presents an opportunity for platforms to take an active role in protecting the societies they operate in – safeguarding public discourse from threats posed by unknown actors who exploit their systems for manipulation and harm.

AUTHENTIC HUMAN INTERACTION AND HUMAN RIGHTS

A rights-based approach to digital governance prioritizes human agency over automated manipulation, ensuring that digital spaces remain environments where genuine discourse and open debate between people thrive. Digital interactions should be reliably human. Authentic pseudonyms – which verify users as real individuals without exposing their identities – offer a mechanism for distinguishing between human and artificial activity while preserving privacy and freedom of expression.[93] Unlike synthetic forces, individuals possess moral reasoning, intentionality, and legal personhood, all of which underpin the legitimacy of human rights protections.[94] A system that prioritizes human interaction and excludes inauthentic actors would reinforce the foundational principles of democratic deliberation and organic communication, ensuring that opinion formation remains free from AI-driven distortion and artificial manipulation.

Germany's ID system, in operation since 2011, offers a practical model for authentic pseudonymous identities, enabling users to generate cryptographically secure tokens for digital participation.[95] This approach balances privacy with accountability, allowing individuals to control their personal data while ensuring trustworthy verification. By issuing pseudonymous tokens that cannot be linked to individuals, the system minimizes risks associated with centralized data storage, such as breaches or misuse. Its decentralized and privacy-preserving design demonstrates the feasibility of scaling such frameworks globally, including in diverse legal, cultural, and technological contexts.

This system proposes establishing a digital identity framework that enables users to interact pseudonymously while ensuring that bots and automated accounts are clearly identified.[96] By ensuring that platforms verify the identities behind automated accounts, this approach holds digital intermediaries – such as social media platforms, identity verification providers, and regulated digital service operators – accountable for preventing manipulation, ensuring transparency, and safeguarding online systems.

Ultimately, removing synthetic engagement and reinforcing human interaction in digital spaces is not merely a technical fix; it is a structural reform necessary to safeguard the human right to freedom of expression and opinion. Disinfo-ops erode these rights by overwhelming real voices and manipulating opinion formation. By prioritizing human agency over artificial distortion and ensuring privacy-preserving verification measures, societies can restore the conditions for authentic speech, reclaim digital spaces for real human engagement, and protect the foundations of democracy, knowledge production, and human rights.

CONCLUSIONS AND RECOMMENDATIONS

The rise of synthetic forces has magnified the scale, sophistication, and targeting accuracy of disinformation campaigns to unprecedented levels. Addressing the information crisis requires more than piecemeal detection efforts – it demands structural reform. As synthetic activity accelerates at a high velocity, we enter an arms race between advancing technological forces and the pursuit of authenticity, where failing to act decisively now will only deepen the imbalance. Just as environmental protections are essential to clean air and water, preserving a healthy information ecosystem must include identifying and filtering out pollution.

To conclude, three key areas are put forward to strengthen the needed digital integrity: creating genuine human interaction online, ensuring privacy-preserving data access for independent research, and expanding digital literacy initiatives to equip individuals with the skills needed to navigate today's complex information landscape. Taken together, these strategies provide a human rights-aligned foundation for mitigating the harms of digital disinfo-ops and reinforcing trust in the digital sphere and beyond.

AUTHENTIC PSEUDONYMS

The proposal of authentic pseudonyms – where users engage online under non-identifiable aliases linked to verified accounts – offers a middle ground between full anonymity and real-name requirements.[97] This system would allow everyone to maintain their privacy while preventing tech-driven manipulation and large-scale information operations, offering critical protection for whistleblowers, activists, and other vulnerable groups. Crucially, ensuring that humans interact with humans in online discourse upholds the essence of human rights law on freedom of expression and opinion, where a fair competition of ideas fuels debate, scientific progress, and positive societal development.

DATA ACCESS

Ensuring privacy-preserving data access is critical to understanding and countering digital disinformation. Independent researchers require structured access to platform data to analyze the quantified size of disinfo-ops, how false narratives spread, the role of algorithmic amplification, and the impact of disinformation on public discourse. The EU Digital Services Act offers a promising model by mandating researcher access to platform data while preserving the privacy of users – expanding global frameworks for data transparency is essential for strengthening evidence-based policy responses.[98]

DIGITAL LITERACY

Digital literacy is a long-term, foundational solution for strengthening societal resilience against disinformation. It equips individuals with critical thinking skills to assess information, recognize manipulation tactics, and navigate algorithm-driven content environments. Yet, because digital literacy is relevant across all disciplines but owned by none, integrating it into education remains a challenge. Librarians, as experts in information science, are key allies in this effort, helping structure literacy programs that promote accuracy, verification, and responsible information use. Governments, educators, and civil society must collaborate to embed digital literacy into curricula, expand public awareness campaigns, and tailor programs to diverse communities, ensuring accessibility across different linguistic, cultural, and socioeconomic groups.[99]

END NOTES

1 J Haider and O Sundin, Paradoxes of Media and Information Literacy: The Crisis of Information (Taylor & Francis 2022).

2 ibid; S Livingstone, Tackling the Information Crisis: A Policy Framework for Media System Resilience, Report of the LSE Commission on Truth, Trust and Technology (London School of Economics and Political Science 2019) <https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis.pdf> accessed December 2024.

3 Lt Gen I Pacepa and R Rychlak, Disinformation (WND Books, 2013); Ocherki Istorya Rossiiskoy Vneshny Razvedki, Mezhdunarodniye Otnsheniya (1996) vol 2, 13-14 (Y.M. Primakov, ed), cited in H Romerstein, Disinformation as a KGB Weapon in the Cold War, (2001) 1(1) Journal of Intel History 54; J Darczewska, P Żochowski, 'Active Measures, Russia's Key Export' (2017) 12 Point of View (Centre for Eastern Studies, No. 64).

4 I Khan, Report of the Special Rapporteur on Disinformation and Freedom of Opinion and Expression (13 April 2021) UN Doc A/HRC/47/25; United Nations General Assembly (UNGA), Countering Disinformation for the Promotion and Protection of Human Rights and Fundamental Freedoms UNGA Res 76/227 (24 December 2021) UN Doc A/RES/76/227; UN Human Rights Council, Role of States in Countering the Negative Impact of Disinformation on the Enjoyment and Realization of Human Rights UN Doc A/HRC/49/L.31/Rev.1 (30 March 2022).

5 United Nations Group of Governmental Experts (UN GGE), Report on Developments in the Field of Information and Telecommunications in the Context of International Security (24 June 2013) UN Doc A/68/98, para 19; UN GGE Report (22 July 2015) UN Doc A/70/174, paras 24–28; UN Open-ended Working Group, Report on Developments in the Field of Information and Telecommunications in the Context of International Security (2021) UN Doc A/75/816, para 34.

6 S Bradshaw and P Howard, The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation (Project on Computational Propaganda 2019) 11 (my emphasis).

7 C Wardle and H Derakhshan, Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking (Council of Europe 2017) 10 n 6: the term "information pollution" is traced to Jakob Nielsen in 2003.

8 Khan report, Disinformation... (n 4) para 9-15.

9 M Schmitt, 'Grey Zones in the International Law of Cyberspace' (2017) 42(2) Yale Journal of International Law Online 1; M Schmitt, '"Virtual" Disenfranchisement: Cyber Election Meddling in the Grey Zones of International Law' (2018) 19(1) Chicago Journal of International Law 31; A Sari, 'Legal Resilience in an Era of Grey Zone Conflicts and Hybrid Threats' (2020) 33(6) Cambridge Review of International Affairs; M Regan and A Sari (eds), Hybrid Threats and Grey Zone Conflict (OUP 2024).

10 Wardle and Derakshan, Information Disorder (n 7) the concept of mal-information (genuine information used to cause harm) is introduced in their framework, but this report will focus on mis- and disinformation; D Fallis, 'What is Disinformation?' (2015) 63(3) Library Trends 401; C Jack, 'Lexicon of Lies: Terms for Problematic Information' (Data & Society 2017) <https://datasociety.net/library/lexicon-of-lies/> accessed Dec 2024.

11 Wardle and Derakshan, Information Disorder (n 7); UNESCO, Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training (2020) <https://en.unesco.org/fightfakenews> accessed Dec 2024; S Woolley and P Howard, Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media (OUP 2018); K Starbird, A Arif, and T Wilson, 'Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations' (2019) 3 Proceedings of the ACM on Human-Computer Interaction CSCW, Article 127; M Ressa, How to Stand Up to a Dictator (WH Allen 2022); Working Group on Infodemics, Policy Framework (Forum on Information and Democracy, Nov 2020) <https://informationdemocracy.org/wp-content/uploads/2020/11/ForumID_Report-on-infodemics_101120.pdf> accessed Dec 2024.

12 J Kavanagh and MD Rich, Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life (RAND Corporation 2018).

13 See (n 3); SJ Barela and J Duberry, 'Understanding Disinformation Operations in the 21st Century' in JD Ohlin and D Hollis (eds), Defending Democracies: Combating Foreign Election Interference in a Digital Age (OUP 2021).

14 G Curtis, An Overview of Psychological Operations (PSYOP) (Federal Research Division, Library of Congress, October 1989).

15 C Jack, 'Lexicon of Lies' (n 10); NATO, Allied Command Transformation develops the Cognitive Warfare Concept to Combat Disinformation and Defend Against "Cognitive Warfare" (3 July 2024) <https://www.act.nato.int/article/cogwar-concept/> accessed Jan 2024; E Broda and J Strömbäck, 'Misinformation, Disinformation, and Fake News: Lessons from an Interdisciplinary, Systematic Literature Review' (2024) 48 Annals of the International Communication Association 139.

16 For introduction of the term 'disinfo-ops' see SJ Barela and J Duberry (n 13); see also SJ Barela, 'Dawning Digital Data Access via New EU Law' (Just Security, 20 Oct 2022) <https://www.justsecurity.org/83622/dawning-digital-data-access-via-new-eu-law/> accessed Jan 2025.

17 President B Obama, Exec Order No 13757, 3 CFR 391 (2016) (issued in response to Russian interference in the 2016 U.S. presidential election); R Mueller, Report on the Investigation Into Russian Interference in the 2016 Presidential Election (Volume 1) (US Department of Justice 2019) 14-35.

18 Y Benkler, R Faris, and H Roberts, Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics (OUP 2018) chs 4-6; B Adair, Beyond the Big Lie: The Epidemic of Political Lying, Why Republicans Do It More, and How It Could Burn Down Our Democracy (Atria Books 2024).

19 See S. Barela, Info-Brief on EU Data Access for Mis- and Disinformation, Geneva Academy of International Humanitarian Law and Human Rights, 2024.

20 For one recent example, see The Times, 'Romania Cancels Election After "Russian Meddling" on TikTok' (6 December 2024) <https://www.thetimes.co.uk/article/romania-cancels-election-and-warns-of-russian-meddling-on-tiktok-z0khnw3m7> accessed Dec 2024; see also FBI and CISA Issue Public Service Announcement Warning of Tactics Foreign Threat Actors Are Using (18 October 2024) <https://www.cisa.gov/news-events/news/fbi-and-cisa-issue-public-service-announcement-warning-tactics-foreign-threat-actors-are-using> accessed Dec 2024; R Ó Fathaigh, T Dobber, F Zuiderveen Borgesius, and J Shires, "Microtargeted Propaganda by Foreign Actors: An interdisciplinary exploration" Maastricht Journal of European and Comparative Law 2021, Vol. 28(6) 856–877 (2021).

21 K Starbird and T Wilson, 'Cross-Platform Disinformation Campaigns: Lessons Learned, Next Steps' (2020) 1(1) Harvard Kennedy School Misinformation Review <https://doi.org/10.37016/mr-2020-002> accessed Dec 2024.

22 Barela, 'Dawning Digital Data Access' (n 16).

23 ibid; Wardle and Derakshan, Information Disorder (n 7); see generally P Howard, Lie Machines: How to Save Democracy from Troll Armies, Deceptive Algorithms, and Fake News (Yale University Press 2020).

24 For a proposal to curb automated and synthetic activity, preserve privacy and promote human interaction, see S Hallensleben, Trust in the European Digital Space in the Age of Automated Bots and Fakes (European Observatory of ICT Standardisation, January 2022) <https://www.standict.eu/news/trusted-information-digital-space> accessed Dec 2024.

25 J Desjardins, (2019) How Much Data Is Generated Each Day? World Economic Forum (17 April 2019) <https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/> accessed Dec 2024; Office of the High Commissioner of Human Rights, The Right to Privacy in the Digital Age: Report of the United Nations High Commissioner for Human Rights, (3 August 2018) A/

HRC/39/29.

26 J Chester and K Montgomery, 'The Role of Digital Marketing in Political Campaigns' (2017) 6 International Political Review; K Endres and KJ Kelly, 'Does Microtargeting Matter? Campaign Contact Strategies and Young Voters' (2018) 28 Journal of Elections, Public Opinion and Parties 1–18.

27 ibid, Chester and Montgomery, "Role of Digital Marketing…"; Z Tufekci, 'Engineering the Public: Big Data, Surveillance and Computational Politics' (2014) 19 First Monday; Fathaigh et al. "Microtargeted…" (n 20).

28 Privacy International, Written Evidence Submitted to the Joint Committee on Human Rights Inquiry into Privacy and Data Rights (28 March 2019) paras 3.4-3.6 <https://committees.parliament.uk/writtenevidence/100839/html/> accessed Jan 2025.

29 Howard, Lie Machines (n 23) ch 1; Almog Simchon, Matthew Edwards, and Stephan Lewandowsky, 'The Persuasive Effects of Political Microtargeting in the Age of Generative Artificial Intelligence' (2024) 3 PNAS Nexus <https://pubmed.ncbi.nlm.nih.gov/38328785/> accessed Dec 2024. Cf Jessica Baldwin-Philippi, 'The Myths of Data-Driven Campaigning' (2017) 34 Political Communication 627–633.

30 ibid, Howard, Lie Machines (n 23) ch 3; see generally, Woolley and Howard, Computational Propaganda (n 10); NATO Strategic Communications Centre of Excellence, 'Robotrolling' (2021); S Vosoughi, D Roy, and S Aral, 'The Spread of True and False News Online' (2018) 359 Science 1146–1151.

31 ibid, Howard, Lie Machines (n 23) ch 2.

32 L Ng, D Robertson, and K Carley, 'Cyborgs for Strategic Communication on Social Media' (2024) 1 Big Data & Society.

33 See S. Barela, Part II – Charting the International Legal Frameworks, Geneva Academy of International Humanitarian Law and Human Rights, 2024.

34 For granular detail on one troll farm see (Mueller) Indictments, US v Internet Research Agency et al No 1:18-cr-32-DLF (D DC, 16 February 2018); Woolley & Howard, Computational Propaganda (n 11); Bradshaw & Howard, "The Global Disinformation Order…' (n 6); D Linvill and P Warren, 'Troll Factories: Manufacturing Specialized Disinformation on Twitter' (2020) Political Communication 1–21; Howard, Lie Machines (n 23) 12: "By 2020, seven countries were running misinformation campaigns targeting citizens in other countries: along with Russia and China, there were similar operations in India, Iran, Pakistan, Saudia Arabia, and Venezuela".

35 L Gaur (ed), DeepFakes: Creation, Detection, and Impact (1st edn, CRC Press 2022); C Doss et al., 'Deepfakes and Scientific Knowledge Dissemination' (2023) 13(1) Nature Portfolio, Scientific Reports 13429 <https://doi.org/10.1038/s41598-023-39944-3> accessed Dec 2024; B Jacobsen and J Simpson, 'The Tensions of Deepfakes' (2024) 27(6) Information, Communication & Society 1095–1109 <https://doi.org/10.1080/1369118X.2023.2234980> accessed Dec 2024.

36 L Belli & M Wisniak, 'What's in an Algorithm? Empowering Users Through Nutrition Labels for Social Media Recommender Systems' (2023) Knight First Amendment Institute <https://knightcolumbia.org/content/whats-in-an-algorithm-empowering-users-through-nutrition-labels-for-social-media-recommender-systems> accessed Dec 2024.

37 Wardle and Derakhshan, Information Disorder (n 7) 49-56; M Cinelli, A Galeazzi, W Quattrociocchi & M Starnini, 'The Echo Chamber Effect on Social Media' (2021) 118(9) Proceedings of the National Academy of Sciences e2023301118 <https://doi.org/10.1073/pnas.2023301118> accessed Dec 2024.

38 See Howard, Lie Machines (n 23) at 25. For a proposal to mitigate the problem see, Belli & Wisniak, 'What's in an Algorithm?' (n 36).

39 Howard, Lie Machines (n 23) iv.

40 Scholars have explored the difficulty to study the vulnerability of women in a low-literacy marginalized population in rural India, see A Bhattacharya, E Jorgensen, U Naik, and CT Moore, 'Measuring Susceptibility to Misinformation in Lower-Income Populations' (15 August 2024) The Mercury Project, Social Science Research Council <https://www.ssrc.org/mercury-project/2024/08/15/measuring-susceptibility-to-misinformation-in-lower-income-populations/> accessed Dec 2024.

41 E Michael et al, Decoding Technology-Facilitated Gender-Based Violence: A Reality Check from Seven Countries (Rutgers, 2024) <https://rutgers.international/resources/decoding-technology-facilitated-gender-based-violence-a-reality-check-from-seven-countries/>; L Di Meco, Monetizing Misogyny: Gendered Disinformation and the Undermining of Women's Rights and Democracy Globally (ShePersisted, 2023) <https://she-persisted.org/wp-content/uploads/2023/02/ShePersisted_MonetizingMisogyny.pdf> both accessed Dec 2024.

42 I Khan, Report of the Special Rapporteur on Gendered Disinformation and the Right to Freedom of Opinion and Expression (7 August 2023) UN Doc A/78/288.

43 Meta, 'Removing Myanmar Military Officials From Facebook' (28 August 2018) <https://about.fb.com/news/2018/08/removing-myanmar-officials/> accessed Dec 2024.

44 Independent Investigative Mechanism for Myanmar (IIMM), 'Anti-Rohingya Hate Speech on Facebook: Content and Network Analysis' (27 March 2024) <https://www.legal-tools.org/doc/0nji9u/>.

45 A second report released by the IIMM further demonstrates the plight of marginalized communities. It examines Myanmar state authorities' severely lacking response to sexual and gender-based crimes committed by security forces against the Rohingya women during the 2016 and 2017 clearance operations: IIMM, 'Efforts to Investigate and Punish Sexual and Gender-Based Crimes Committed Against Rohingya: Evidence Analysis' (27 March 2024) <https://www.legal-tools.org/doc/0kbf7f/)> accessed Dec 2024.

46 See eg S Khakimullin, 'Climate disinformation' (European Commission, March 2024) <https://climate.ec.europa.eu/eu-action/climate-disinformation_en> accessed Jan 2025; for challenges to scientific knowledge generally, see J Haider, K Rolf Söderström, B Ekström, and M Rödl, 'GPT-fabricated scientific papers on Google Scholar: Key features, spread, and implications for preempting evidence manipulation' (2024) 5(5) Harvard Kennedy School Misinformation Review; Doss et al., 'Deepfakes and Scientific Knowledge Dissemination'(n 35).

47 "But we're not just fighting an epidemic; we're fighting an infodemic": WHO Director-General Tedros Adhanom Ghebreyesus, 'Speech at the Munich Security Conference' (15 February 2020) <https://www.who.int/director-general/speeches/detail/munich-security-conference> accessed Dec 2024.

48 World Health Organization, 'Let's flatten the infodemic curve' <https://www.who.int/news-room/spotlight/let-s-flatten-the-infodemic-curve>; see also World Health Organization on Infodemic <https://www.who.int/health-topics/infodemic#tab=tab_1>; M Richtel "W.H.O. Fights a Pandemic Besides Coronavirus: An 'Infodemic'" New York Times (6 Feb 2020) <https://www.nytimes.com/2020/02/06/health/coronavirus-misinformation-social-media.html>.

49 K Singh et al., 'Misinformation, believability, and vaccine acceptance over 40 countries: Takeaways from the initial phase of the COVID-19 infodemic' (2021) 16(6) PLOS ONE e0263381 <https://doi.org/10.1371/journal.pone.0263381> accessed Jan 2025; MS Islam et al., 'COVID-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation' (2021) 16(5) PLOS ONE e0251605 <https://doi.org/10.1371/journal.pone.0251605> accessed Jan 2025.

50 C Radsch, 'AI and Disinformation: State-Aligned Information Operations and the Distortion of the Public Sphere' (OSCE Representative on Freedom of the Media, 2022) <https://www.osce.org/fom/sofjo> accessed Dec 2024.

51 See eg the response from France Soir, « NewsGuard continue la désinformation sélective et ciblée, fake news et censure » (10 août 2020) < https://www.francesoir.fr/societe-science-tech/newsguard-continue-la-desinformation-selective-et-ciblee-fake-news-et-censure> accessed Dec 2024.

52 Le Monde avec AFP, « France-Soir » perd son agrément de service de presse en ligne », (21 August 2024) <https://www.lemonde.fr/actualite-medias/

article/2024/08/21/france-soir-perd-son-agrement-de-service-de-presse-en-ligne_6289648_3236.html> accessed Dec 2024. Relatedly, in early 2024, a large network spreading pro-Russian propaganda was exposed by cyber experts in France, Radio France Internationale (RFI), 'French cyber experts reveal vast network of Russian disinformation sites' (13 Feb 2024) <https://www.rfi.fr/en/international/20240213-french-cyber-experts-reveal-vast-network-of-russian-disinformation-sites> accessed Dec 2024.

53 SJ Barela, 'Cross-Border Cyber Ops to Erode Legitimacy: An Act of Coercion' (Just Security, 12 January 2017) <https://www.justsecurity.org/36212/cross-border-cyber-ops-erode-legitimacy-act-coercion/>; SJ Barela, 'Bots, Trolls and Dezinformatsiya: The Continuing Russian Campaign to Divide the Democratic Party in the USA' (E-International Relations, 27 Sept 2017) <https://www.e-ir.info/2017/09/27/the-continuing-russian-campaign-to-divide-the-democratic-party-in-the-usa/>; SJ Barela, 'Zero Shades of Grey: Russian-Ops Violate International Law' (Just Security, 29 March 2018) <https://www.justsecurity.org/54340/shades-grey-russian-ops-violate-international-law/>; SJ Barela, 'Disobeying Trump: "Context and Consequences" of Russian Ops', (Just Security, 27 Sept 2018) <https://www.justsecurity.org/60883/disobeying-trump-context-consequences-russian-ops/> accessed Dec 2024.

54 Mueller, 'Report on Russian Interference…' (n 17) 14–35; US v Internet Research Agency et al (n 34); United States Select Committee on Intelligence, US Senate, Russian Active Measures Campaigns and Interference in the 2016 US Election. Volume 1: Russian Efforts Against Election Infrastructure (116th Congress 2019–2020, 1st Session, Report 116-XX); United States Select Committee on Intelligence, United States Senate, Russian Active Measures Campaigns and Interference in the 2016 US Election. Volume 2: Russia's Use of Social Media (116th Congress 2019–2020, 1st Session, Report 116-XX).

55 US House Select Committee to Investigate the January 6th Attack on the United States Capitol, Final Report of the Select Committee to Investigate the January 6th Attack on the United States Capitol (US Government Publishing Office 2022).

56 Brazilian Congressional Inquiry, Draft Report on the January 8 Insurrection: Recommendations for Criminal Charges Against Jair Bolsonaro and Others (October 2023); Brazilian Federal Police, Report on the Indictment of Jair Bolsonaro and Others for the January 8 Attacks on Democratic Institutions (November 2024); J Ozawa, J Lukito, F Bailez, and L Fakhouri, 'Brazilian Capitol Attack: The Interaction Between Bolsonaro's Supporters' Content, WhatsApp, Twitter, and News Media' (2024) 5(2) Harvard Kennedy School Misinformation Review.

57 F Fukuyama, Trust: The Social Virtues and the Creation of Prosperity (Free Press 1995).

58 Z Tufekci, 'Engineering the Public…' (n 27).

59 Wardle and Derakhshan, Information Disorder (n 7); UNESCO, 'Journalism, Fake News, and Disinformation' (n 11); Starbird, Arif and Wilson, 'Disinformation as Collaborative Work' (n 11); Starbird and Wilson 'Cross-Platform Disinformation Campaigns' (n 21); Radsch, 'AI and Disinformation' (n 50); S Vosoughi, D Roy, and S Aral, 'The Spread of True and False News Online' (2018) 359 Science 1146–1151.

60 For a proposal to curb automated and synthetic activity, preserve privacy and promote human interaction online see Hallensleben, Trust in the European Digital Space (n 24).

61 International Covenant on Civil and Political Rights (ICCPR) (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171, art 19.

62 Information Commissioner's Office, Democracy Disrupted? Personal Information and Political Influence (11 July 2018) <https://ico.org.uk/media/2259369/democracy-disrupted-110718.pdf> accessed Jan 2025.

63 Privacy International, Written Evidence Submitted… (n 28) paras 3.4-3.6.

64 Office of the United Nations High Commissioner for Human Rights (OHCHR), The Right to Privacy in the Digital Age, A/HRC/51/17 (4 August 2022) para 2.

65 Universal Declaration of Human Rights (adopted 10 December 1948) UNGA Res 217 A(III), art 19.

66 ICCPR, art 19.

67 ibid, art 19(3).

68 UN Human Rights Committee (HRC), General comment No 34, Article 19, Freedoms of opinion and expression, CCPR/C/GC/34 (12 September 2011) para 22; UN Special Rapporteur on Freedom of Opinion and Expression, OSCE Representative on Freedom of the Media, OAS Special Rapporteur on Freedom of Expression, and ACHPR Special Rapporteur on Freedom of Expression and Access to Information, Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda (3 March 2017) FOM.GAL/3/17 <https://www.osce.org/fom/302796> accessed Dec 2024. On proportionality, see HRC, CCPR General Comment No. 27: Article 12 (Freedom of Movement), CCPR/C/21/Rev.1/Add.9 (2 November 1999) para 14.

69 ICCPR, (n 61) art 20. HRC, 'General Comment No 11: Prohibition of Propaganda for War and Inciting National, Racial or Religious Hatred (Art. 20)' (29 July 1983). It is also prohibited by the UN International Convention on the Elimination of All Forms of Racial Discrimination, adopted 21 December 1965, entered into force 4 January 1969, 660 U.N.T.S. 195.

70 See M Nowak, U.N. Covenant on Civil and Political Rights: CCPR Commentary 441 (1993) at 441–42.

71 E Aswad, 'Losing the Freedom to Be Human' (2020) 52 Columbia Human Rights Law Review 306.

72 ICCPR, (n 61) art 18(1): "Everyone shall have the right to freedom of thought, conscience and religion".

73 Aswad, 'Losing the Freedom to Be Human' (n 71) 353.

74 S Alegre, "Rethinking Freedom of Thought for the 21st Century" European Human Rights Law Review Issue 3 (2017).

75 Khan report, Disinformation… (above n 4) para 66.

76 For a proposal to curb automated and synthetic activity, preserve privacy and promote human interaction online see Hallensleben, Trust in the European Digital Space (n 24).

77 Human Rights Council, General Comment No 34, para 11 (above n 68): "The scope of paragraph 2 embraces even expression that may be regarded as deeply offensive".

78 ibid, para 49: "The Covenant does not permit general prohibition of expressions of an erroneous opinion or an incorrect interpretation of past events".

79 Khan report, Disinformation… (above n 4) para 50-51.

80 HRC, General Comment No 34 (above n 68) para 34.

81 Khan report, Disinformation… (above n 4) para 52-55.

82 See (n 68).

83 N Law et al., 'A Global Framework of Reference on Digital Literacy Skills for Indicator 4.4.2' (UNESCO Institute for Statistics 2018) Information Paper No. 51.

84 Khan report, Disinformation… (n 4) para. 38, 46, 94. At para 23: "There is clear evidence that robust public information regimes and independent journalism are strong antidotes to disinformation."

85 United Nations Human Rights Council, 'Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework' (16 June 2011) UN Doc A/HRC/17/31.

86 D Kaye, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (22 May 2015) UN Doc A/HRC/29/32, paras 27-28, 62; D Kaye, Report on the Role of Digital Technologies in Protecting Freedom of Expression (6 April 2018) UN Doc A/HRC/38/35, paras 9–12; D Kaye, Report on Threats to Freedom of Expression in the Digital Age (9 October 2019) UN Doc A/74/486, paras 40–55; Khan report, Disinformation... (above n 4) paras 63-82 and 95-103; I Khan, Report of the Special Rapporteur on Challenges to Freedom of Expression in the Digital Age (10 August 2022) UN Doc A/77/288, paras 74-99 and 123-131; I Khan, Report on Gendered Disinformation... (n 42) paras 62,64,69.

87 ibid, Khan report, Disinformation, para 3: "Companies play a major role in spreading disinformation but their efforts to address the problem have been woefully inadequate."

88 ibid, para 16: "False information is amplified by algorithms and business models that are designed to promote sensational content that keep users engaged on platforms."

89 Wardle and Derakhshan, Information Disorder (n 7) 49-56.

90 See OHCHR (n 64).

91 Kaye, A/HRC/38/35 (above n 86), Note by the Secretariat: "At a minimum, companies and States should pursue radically improved transparency, from rule-making to enforcement of the rules, to ensure user autonomy as individuals increasingly exercise fundamental rights online."

92 Khan report, Disinformation... (above n 7) para 81: "Most of the largest social media companies produce transparency reports twice a year, but they do not share more precise and meaningful information about action taken to address disinformation or misinformation." See also, Working Group on Infodemics, Policy Framework (above n 11).

93 For a proposal to curb automated and synthetic activity, preserve privacy and promote human interaction online see Hallensleben, Trust in the European Digital Space (n 24).

94 L Higby, 'Navigating the Speech Rights of Autonomous Robots in a Sea of Legal Uncertainty' (2021) 26(1) Journal of Technology Law & Policy at 12-18.

95 J Bender, Ö Dagdelen, M Fischlin, and D Kügler, 'Domain-Specific Pseudonymous Signatures for the German Identity Card' (German Federal Office for Information Security and Darmstadt University of Technology, 2010) <https://eprint.iacr.org/2012/558> accessed Dec 2024.

96 See also, S Adler et al. Artificial Intelligence and the Value of Privacy-Preserving Tools to Distinguish Who is Real Online (OpenAI and collaborators, 2024) <https://arxiv.org/pdf/2408.07892> accessed Dec 2024.

97 For a proposal to curb automated and synthetic activity, preserve privacy and promote human interaction online see Hallensleben, Trust in the European Digital Space (n 24).

98 See further, S. Barela, Info-Brief on EU Data Access, Geneva Academy of International Humanitarian Law and Human Rights (2025).

99 For in-depth critical analysis, see Haider and Sundin (n 1).

THE GENEVA ACADEMY

The Geneva Academy provides post-graduate education, conducts academic legal research and policy studies, and organizes training courses and expert meetings. We concentrate on branches of international law that relate to situations of armed conflict, protracted violence, and protection of human rights.

DISCLAIMER

The Geneva Academy of International Humanitarian Law and Human Rights is an independent academic centre. Our publications seek to provide insights, analysis and recommendations, based on open and primary sources, to policymakers, researchers, media, the private sector and the interested public. The designations and presentation of materials used, including their respective citations, do not imply the expression of any opinion on the part of the Geneva Academy concerning the legal status of any country, territory, or area or of its authorities, or concerning the delimitation of its boundaries. The views expressed in this publication represent those of the authors and not necessarily those of the Geneva Academy, its donors, parent institutions, the board or those who have provided input or participated in peer review. The Geneva Academy welcomes the consideration of a wide range of perspectives in pursuing a well-informed debate on critical policies, issues and developments in international human rights and humanitarian law.

The Geneva Academy of International Humanitarian Law and Human Rights

Villa Moynier
Rue de Lausanne 120B
CP 1063 - 1211 Geneva 1 - Switzerland
Phone: +41 (22) 908 44 83
Email: info@geneva-academy.ch
www.geneva-academy.ch

© The Geneva Academy of International Humanitarian Law and Human Rights

This work is licensed for use under a Creative Commons Attribution-Non-Commercial-Share Alike 4.0 International License (CC BY-NC-ND 4.0).