

Our Common Agenda
Policy Brief 8

**Information
Integrity
on Digital
Platforms**

JUNE 2023



**United
Nations**



Introduction

CHAPEAU

The challenges that we are facing can be addressed only through stronger international cooperation. The Summit of the Future, to be held in 2024, is an opportunity to agree on multilateral solutions for a better tomorrow, strengthening global governance for both present and future generations (General Assembly resolution [76/307](#)). In my capacity as Secretary-General, I have been invited to provide inputs to the preparations for the Summit in the form of action-oriented recommendations, building on the proposals contained in my report entitled “Our Common Agenda” ([A/75/982](#)), which was itself a response to the declaration on the commemoration of the seventy-fifth anniversary of the United Nations (General Assembly resolution [75/1](#)). The present policy brief is one such input. It serves to elaborate on the ideas first proposed in Our Common Agenda, taking into account subsequent guidance from Member States and more than one year of intergovernmental and multi-stakeholder consultations, and rooted in the purposes and the principles of the Charter of the United Nations, the Universal Declaration of Human Rights and other international instruments.

PURPOSE OF THIS POLICY BRIEF

The present policy brief is focused on how threats to information integrity are having an impact on progress on global, national and local issues. In Our Common Agenda, I called for empirically backed consensus around facts, science and knowledge. To that end, the present brief outlines potential principles for a code of conduct that will help to guide Member States, the digital platforms and other stakeholders in their efforts to make the digital space more inclusive and safe for all, while vigorously defending the right to freedom of opinion and expression, and the right to access information. The Code of Conduct for Information Integrity on Digital Platforms is being developed in the context of preparations for the Summit of the Future. My hope is that it will provide a gold standard for guiding action to strengthen information integrity.

Digital platforms are crucial tools that have transformed social, cultural and political interactions everywhere. Across the world, they connect concerned global citizens on issues that matter. Platforms help the United Nations to inform and engage people directly as we strive for peace, dignity and equality on a healthy planet. They have given people hope in times of crisis and struggle, amplified voices that were previously unheard and breathed life into global movements.

Yet these same platforms have also exposed a darker side of the digital ecosystem. They have enabled the rapid spread of lies and hate, causing real harm on a global scale. Optimism over the potential of social media to connect and engage people has been dampened as mis- and disinformation and hate speech have surged from the margins of digital space into the mainstream. The danger cannot be overstated. Social media-enabled hate speech and disinformation can lead to violence and death.¹ The ability to disseminate large-scale disinformation to undermine

scientifically established facts poses an existential risk to humanity ([A/75/982](#), para. 26) and endangers democratic institutions and fundamental human rights. These risks have further intensified because of rapid advancements in technology, such as generative artificial intelligence. Across the world, the United Nations is monitoring how mis- and disinformation and hate speech can threaten progress towards the Sustainable Development Goals. It has become clear that business as usual is not an option.

IN PRACTICE, THE DISTINCTION
BETWEEN MIS- AND DISINFORMATION

CAN BE SUBTLE AND

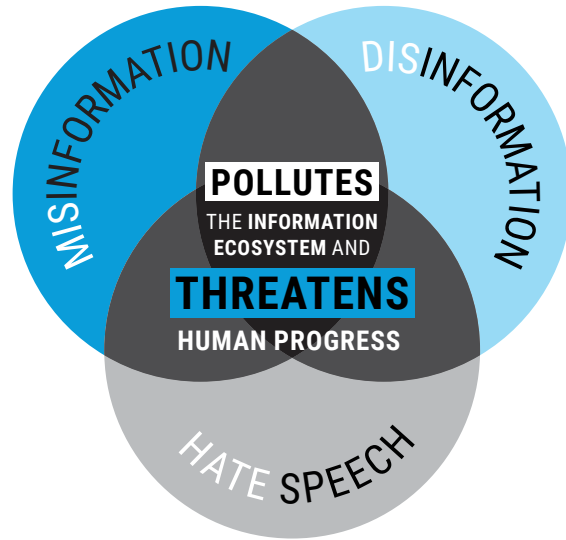
DIFFICULT TO

DETERMINE

What is information integrity?

Information integrity refers to the accuracy, consistency and reliability of information. It is threatened by disinformation, misinformation and hate speech. While there are no universally accepted definitions of these terms, United Nations entities have developed working definitions.

The Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression refers to disinformation as “false information that is disseminated intentionally to cause serious social harm”.² Disinformation is described by the United Nations Educational, Scientific and Cultural Organization (UNESCO) as false or misleading content that can cause specific harm, irrespective of motivations, awareness or behaviours.³



For the purposes of the present policy brief, the difference between mis- and disinformation lies with intent. Disinformation is information that is not only inaccurate, but is also intended to deceive and is spread in order to inflict harm. Disinformation can be spread by State or non-State actors in multiple contexts, including during armed conflict, and can affect all areas of development, from peace and security to human rights, public health, humanitarian aid and climate action.

Misinformation refers to the unintentional spread of inaccurate information shared in good faith by those unaware that they are passing on falsehoods. Misinformation can be rooted in disinformation as deliberate lies and misleading narratives are weaponized over time, fed into the public discourse and passed on unwittingly.⁴ In practice, the distinction between mis- and disinformation can be difficult to determine.

Hate speech, according to the working definition in the United Nations Strategy and Plan of Action on 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”.⁵

Mis- and disinformation and hate speech are related but distinct phenomena, with certain areas of overlap and difference in how they can be identified, mitigated and addressed. All three pollute the information ecosystem and threaten human progress.⁶

Threats to information integrity are not new. Falsehoods and hatred have long been spread for political or financial gain. Yet in the digital age these operations can be conducted on a previously unthinkable scale. Powerful communication tools can now spread content instantly across the globe, creating a problem so widespread that online platforms themselves are at times unable to grasp its full extent. The lack of governmentally agreed definitions of these terms should not result in inertia. We must do all we can to curb the harms they cause.

Information integrity and digital platforms

Digital platforms should be integral players in the drive to uphold information integrity. While certain traditional media can also be sources of mis- and disinformation, the velocity, volume and virality of their spread via digital channels warrants an urgent and tailored response. For the purposes of the present brief, the term “digital platform” refers to a digital service that facilitates interactions between two or more users, covering a wide range of activities, from social media and search engines to messaging apps. Typically, they collect data about their users and their interactions.⁷

Mis- and disinformation are created by a wide range of actors, with various motivations, who by and large are able to remain anonymous. Coordinated disinformation campaigns by State and non-State actors have exploited flawed digital systems to promote harmful narratives, with serious repercussions.

Many States have launched initiatives to regulate digital platforms, with at least 70 such laws adopted or considered in the last four years.⁸ At their core, legislative approaches typically involve a narrow scope of remedies to define and remove harmful content. By focusing on the removal of harmful content, some States have introduced flawed and overbroad legislation that has in effect silenced “protected speech”, which is permitted under international law. Other responses, such as blanket Internet shutdowns and bans on platforms, may lack legal basis and infringe human rights.

Many States and political figures have used alleged concerns over information integrity as a pretext to restrict access to information, to discredit and restrict reporting, and to target journalists and opponents.⁹ State actors have also pressured platforms to do their bidding under the guise of tackling mis- and disinformation.¹⁰ Freedom of expression experts have stressed that State actors have a particular duty in this context and “should not make, sponsor, encourage or further false information” (A/77/287, para. 45).

The risks inherent in the regulation of expression require a carefully tailored approach that complies with the requirements of legality, necessity and proportionality under human rights law, even when there is a legitimate public interest purpose (ibid., para. 42).

Disinformation is also big business. Both “dark” and mainstream public relations firms, contracted by States, political figures and the private sector, are key sources of false and misleading content.¹¹ One tactic, among others, has been to publish content to fake cloned versions of news sites to make articles seem like they are from legitimate sources.¹² This shadowy business is extremely difficult to track and research so that the true scale of the problem is unknown. Individuals, too, spread false claims to peddle products or services for profit, often targeting vulnerable groups during times of crisis or insecurity.

A dominant approach in the current business models of most digital platforms hinges on the “attention economy”. Algorithms are designed

to prioritize content that keeps users' attention, thereby maximizing engagement and advertising revenue. Inaccurate and hateful content designed to polarize users and generate strong emotions is often that which generates the most engagement, with the result that algorithms have been known to reward and amplify mis- and disinformation and hate speech.¹³

Facing a decline in advertising revenue, digital platforms are seeking alternative avenues for profit beyond the attention economy. For example, paid verification plans, whereby accounts can buy a seal of approval previously used to denote authenticity, have raised serious concerns for information integrity given the potential for abuse by disinformation actors.¹⁴

BILLIONS

USE SOCIAL MEDIA

Source: Kepios, Digital 2023: Global Overview Report (2023)


TWITTER
0.6B


FACEBOOK MESSENGER
0.9B



FACEBOOK

3.0B


SNAPCHAT
0.6B


DOUYIN
0.7B


KUAISHOU
0.6B


TIKTOK
1.1B


WHATSAPP
2.0B


PINTEREST
0.4B


SINO WEIBO
0.6B


WECHAT
1.3B


TELEGRAM
0.7B


YOUTUBE
2.5B


QQ
0.6B


INSTAGRAM
2.0B

What is the relevant international legal framework?

The promotion of information integrity must be fully grounded in the pertinent international norms and standards, including human rights law and the principles of sovereignty and non-intervention in domestic affairs. In August 2022, I transmitted to the General Assembly a report entitled “Countering disinformation for the promotion and protection of human rights and fundamental freedoms”.¹⁵ In the report, I laid out the international human rights law that applies to disinformation, including the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights. Under these international legal instruments, everyone has the right to freedom of expression.¹⁶

Article 19 of the Universal Declaration of Human Rights and article 19 (2) of the Covenant protect the right to freedom of expression, including the freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, and through any media. The human right to freedom of expression is not limited to favourably received information (A/77/287, para. 13). Linked to freedom of expression, freedom of information is itself a right. The General Assembly has stated: “Freedom of information is a fundamental human right and is the touchstone of all the freedoms to which the United Nations is consecrated” (ibid., para. 14). Freedom of expression and access to information may be subject to certain restrictions that meet specific criteria laid out in article 19 (3) of the Covenant.¹⁷ States cannot add additional grounds or restrict expression beyond what is permissible under international law.

The Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to hostility, discrimination or violence, adopted in 2012, provides practical legal and policy guidance to States on how best to implement article 20 (2) of the Covenant and article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination, which prohibit certain forms of hate speech. The Rabat Plan of Action has already been utilized by Member States in different contexts.¹⁸

Hate speech has been a precursor to atrocity crimes, including genocide. The 1948 Convention on the Prevention and Punishment of the Crime of Genocide prohibits “direct and public incitement to commit genocide”.

In its resolution [76/227](#), adopted in 2021, the General Assembly emphasized that all forms of disinformation can negatively impact the enjoyment of human rights and fundamental freedoms, as well as the attainment of the Sustainable Development Goals. Similarly, in its resolution [49/21](#), adopted in 2022, the Human Rights Council affirmed that disinformation can negatively affect the enjoyment and realization of all human rights.

HATE SPEECH
HAS BEEN A PRECURSOR TO
ATROCITY CRIMES
INCLUDING
GENOCIDE



75%

OF UN PEACEKEEPERS SAID

**MIS- AND
DISINFORMATION**

IMPACTED THEIR

SAFETY AND

SECURITY

UN 2022 INTERNAL SURVEY

What harm is being caused by online mis- and disinformation and hate speech?

Online mis- and disinformation and hate speech pose serious concerns for the global public. One study of survey data from respondents in 142 countries found that 58.5 per cent of regular Internet and social media users worldwide are concerned about encountering misinformation online, with young people and people in low-income tiers feeling significantly more vulnerable.¹⁹ Today's youth are digital natives who are more likely to be connected online than the rest of the population, making them the most digitally connected generation in history.²⁰ Around the world, a child goes online for the first time every half second, putting them at risk of exposure to online hate speech and harm, in some cases affecting their mental health.²¹

The impacts of online mis- and disinformation and hate speech can be seen across the world, including in the areas of health, climate action, democracy and elections, gender equality, security and humanitarian response. Information pollution was identified as a significant concern by 75 per cent of United Nations Development Programme country offices in a 2021 survey. It

has severe implications for trust, safety, democracy and sustainable development, as found in a recent UNESCO-commissioned review of more than 800 academic, civil society, journalistic and corporate documents.²²

Mis- and disinformation can be dangerous and potentially deadly, especially in times of crisis, emergency or conflict. During the coronavirus disease (COVID-19) pandemic, a deluge of mis- and disinformation about the virus, public health measures and vaccines began to circulate online.²³ Certain actors exploited the confusion for their own objectives, with anti-vaccine campaigners driving users to sites selling fake cures or preventive measures.²⁴ Many victims of COVID-19 refused to get vaccinated or take basic health precautions after being exposed to mis- and disinformation online.²⁵

Disinformation can likewise prove deadly in already volatile societal and political contexts. In a 2022 report, the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression examined the impact of weaponized information in sowing confusion, feeding hate, inciting violence and prolonging conflict.²⁶ In another report issued in 2022, it was found that disinformation can "involve bigotry and hate speech aimed at minorities, women and any so-called 'others', posing threats not only to those directly targeted, but also to inclusion and social cohesion. It can amplify tensions and divisions in times of emergency, crisis, key political moments or armed conflict".²⁷

SURVEY DATA FROM RESPONDENTS IN

142 COUNTRIES FOUND THAT

58.5%

OF REGULAR INTERNET AND SOCIAL MEDIA USERS WORLDWIDE ARE CONCERNED ABOUT ENCOUNTERING

MISINFORMATION ONLINE

Some of the worst impacts of online harms are in contexts neglected by digital platforms, even where platforms enjoy high penetration rates. Countries in the midst of conflict, or with otherwise volatile contexts that are often less lucrative markets, have not been allocated sufficient resources for content moderation or user assistance. While traditional media remain an important source of news for most people in conflict areas, hatred spread on digital platforms has also sparked and fuelled violence.²⁸ Some digital platforms have faced criticism of their role in conflicts, including the ongoing war in Ukraine.²⁹

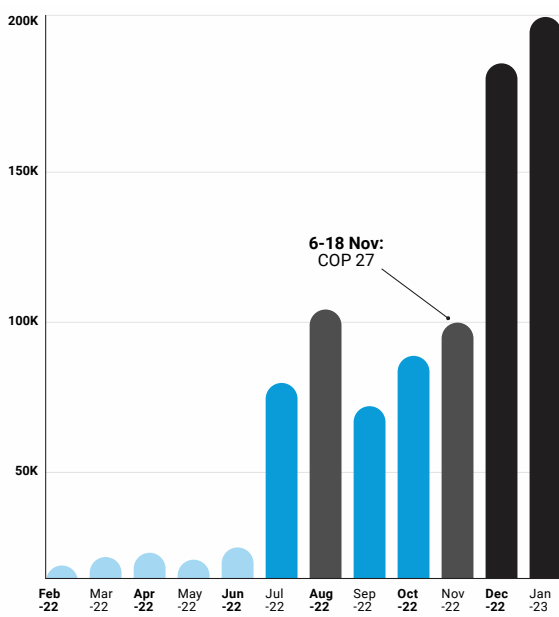
Similarly, mis- and disinformation about the climate emergency are delaying urgently needed action to ensure a liveable future for the planet. Climate mis- and disinformation can be understood as false or misleading content that undercuts the scientifically agreed basis for the existence of human-induced climate change, its causes and impacts. Coordinated campaigns

are seeking to deny, minimize or distract from the Intergovernmental Panel on Climate Change scientific consensus and derail urgent action to meet the goals of the 2015 Paris Agreement. A small but vocal minority of climate science denialists³⁰ continue to reject the consensus position and command an outsized presence on some digital platforms. For example, in 2022, random simulations by civil society organizations revealed that Facebook’s algorithm was recommending climate denialist content at the expense of climate science.³¹ On Twitter, uses of the hashtag #climatescam shot up from fewer than 2,700 a month in the first half of 2022 to 80,000 in July and 199,000 in January 2023. The phrase was also featured by the platform among the top results in the search for “climate”.³² In February 2022, the Intergovernmental Panel on Climate Change called out climate disinformation for the first time, stating that a “deliberate undermining of science” was contributing to “misperceptions of the scientific consensus, uncertainty, disregarded risk and urgency, and dissent”.³³

Some fossil fuel companies commonly deploy a strategy of “greenwashing”, misleading the public into believing that a company or entity is doing more to protect the environment, and less to harm it, than it is. The companies are not acting alone. Efforts to confuse the public and divert attention away from the responsibility of the fossil fuel industry are enabled and supported by advertising and public relations providers, advertising tech companies, news outlets and digital platforms.³⁴ Advertising and public relations firms that create greenwashing content and third parties that distribute it are collectively earning billions from these efforts to shield the fossil fuel industry from scrutiny and accountability. Public relations firms have run hundreds of campaigns for coal, oil and gas companies.³⁵

Mis- and disinformation are having a profound impact on democracy, weakening trust in democratic institutions and independent media, and

FIGURE I
MONTHLY NUMBER OF USES OF #CLIMATESCAM ON TWITTER



Source: Department of Global Communications, using data from Talkwalker

70%

OF UN PEACEKEEPERS SAID

MIS- AND DISINFORMATION

WERE HAVING A

SEVERE, CRITICAL

OR MODERATE IMPACT ON THEIR WORK

dampening participation in political and public affairs. Throughout the electoral cycle, exposure to false and misleading information can rob voters of the chance to make informed choices. The spread of mis- and disinformation can undermine public trust in electoral institutions and the electoral process itself – such as voter registration, polling and results – and potentially result in voter apathy or rejection of credible election results. States and political leaders have proved to be potent sources of disinformation, deliberately and strategically spreading falsehoods to maintain or secure power, or undermine democratic processes in other countries.³⁶

Marginalized and vulnerable groups are also frequent targets of mis- and disinformation and hate speech, resulting in their further social, economic and political exclusion. Women candidates, voters, electoral officials, journalists and civil society representatives are targeted with gendered disinformation online.³⁷ These attacks undermine political participation and weaken democratic institutions and human rights, including the freedom of expression and access to information of these groups.³⁸ This must be an increasingly urgent priority for the international community, not least because more than 2 billion voters are set to go to the polls around the world in 2024.

Mis- and disinformation also cross-pollinate between and within platforms and traditional media, becoming even more complex to track and address if not detected at the source. Disinformation can be a deliberate tactic of ideologically influenced media outlets co-opted by political and corporate

interests.³⁹ At the same time, the rise of digital platforms has precipitated a dramatic decline in trustworthy, independent media. News audiences and advertising revenues have migrated en masse to Internet platforms – a trend exacerbated by the COVID-19 pandemic. “Media extinction” or “news deserts”, where communities lose trustworthy local news sources, can be seen in some regions or countries,⁴⁰ contributing to pollution of the information ecosystem. “Newswashing” – whereby sponsored content is dressed up to look like reported news stories – is often inadequately signposted when posted to digital platforms, lending it a veneer of legitimacy. Once picked up by other media, cited by politicians, or shared widely across platforms, the origin of the information becomes increasingly murky, and news consumers are left unable to distinguish it from objective fact.

Disinformation is also having a direct impact on the work of the United Nations. Resident Coordinators, envoys, mediators and peacekeepers have raised concerns about the effect of disinformation on the Organization’s operational safety, effectiveness and ability to deliver. In a 2022 survey, 70 per cent of United Nations peacekeepers said mis- and disinformation were having a severe, critical or moderate impact on their work, while 75 per cent said that they were having an impact on their safety and security. Mis- and disinformation can also be used to target humanitarian actors and hamper life-saving operations in conflict areas.

FIGURE II

INFORMATION INTEGRITY AND THE SUSTAINABLE DEVELOPMENT GOALS

As shown, threats to information integrity can have a negative impact on the achievement of the Sustainable Development Goals.

	<p>Mis- and disinformation continue to have implications for poverty eradication efforts and the global economy. Economic hardship can also fuel the spread of polarizing and hateful lies, including about marginalized groups. The cost-of-living crisis has been particularly fertile ground for the dissemination of disinformation falsely blaming the switch to renewable energy for soaring energy costs or job losses, for instance.</p>		<p>Mis- and disinformation and hate speech spread online are polarizing societies and targeting already marginalized and vulnerable communities, and can result in their further social, economic and political exclusion.</p>
	<p>Threats to information integrity can compound global hunger, including by exacerbating conflict, climate change, disasters, poverty and inequality. Disinformation can deflect attention and distract from the challenges to global food security posed by conflicts.</p>		<p>Efforts to make cities and communities more sustainable can be undermined by disinformation that denies or deflects attention away from the impacts of human activity on the environment.</p>
	<p>During the coronavirus disease (COVID-19) pandemic, an infodemic of related mis- and disinformation undermined public health measures and vaccination drives. The threat to children's health and well-being resulting from exposure to harmful content persists.</p>		<p>Activists behind initiatives to foster a circular economy and boost zero-waste practices have been targeted through online hate speech and disinformation.</p>
	<p>Mis- and disinformation and hate speech can have an adverse impact on access to quality education, in particular for marginalized groups, including young women and girls. Access to information and digital media literacy drives to increase resilience will play a key role in limiting the societal impact of online harms.</p>		<p>Climate disinformation, and the inertia that it encourages, is undermining efforts to take urgent action to address the climate crisis, including by impeding the crucial shift from polluting fossil fuels to renewable energy and urgent investments in climate resilience.</p>
	<p>Gender-based hate speech and disinformation seek to systematically subjugate women by silencing them and pushing them out of the public sphere. They can have devastating consequences, from suppressing women's voices and fuelling self-censorship, to causing professional and reputational damage and inciting physical violence.</p>		<p>Mis- and disinformation can have a negative impact on efforts to conserve and sustainably use the oceans, seas and marine resources.</p>
	<p>Two billion people live without safely managed drinking water services. Mis- and disinformation about the safety of drinking water and sanitation can have dangerous health consequences.^a</p>		<p>Environmental activists working to protect life on land have been targeted by online hate and disinformation campaigns, with real-life consequences. Climate mis- and disinformation are undermining climate action efforts.^c</p>
	<p>Climate mis- and disinformation, much of it seeded by the fossil fuel industry, are undermining the urgent transition to cleaner forms of energy production, narrowing the closing window to deliver a sustainable future for all.</p>		<p>Disinformation and hate speech have been used to influence elections and public narratives and sow confusion. They have been used to undermine adversaries, thwart peacemaking efforts, incite violence, prolong conflict and damage trust in the rule of law. Efforts to promote peaceful and inclusive societies, and the role of the United Nations in supporting peace and security, have been seriously affected as a result.^d</p>
	<p>Research has pointed to the detrimental impacts of mis- and disinformation and hate speech on the economy.^b</p>		<p>Mis- and disinformation and hate speech can hinder meaningful partnerships to achieve the Goals, while resources diverted to address the problem can weaken efforts to leave no one behind.</p>
	<p>Mis- and disinformation and hate speech, and overbroad responses to these phenomena, can have a detrimental impact on innovation, including by limiting the potential of marginalized groups and making digital spaces less equal and inclusive.</p>		

a World Health Organization and United Nations Children's Fund (UNICEF), *Progress on Household Drinking Water, Sanitation and Hygiene 2000–2020: Five Years into the SDGs* (Geneva, 2021).

b See Roberto Cavazos and CHEQ, "The economic cost of bad actors on the Internet: fake news, 2019"; and London Economics, "The cost of lies: assessing the human and financial impact of COVID-19 related online misinformation on the UK", December 2020.

c Global Witness, *Last Line of Defence: The Industries Causing the Climate Crisis and Attacks against Land and Environmental Defenders* (2021).

d [A/77/288](#).

MONTHS TAKEN TO ADD

100 MILLION

**GLOBAL MONTHLY
ACTIVE USERS**

Source: Similarweb, using data from Sensor Tower



INSTAGRAM

30 MONTHS



TIKTOK

9 MONTHS



CHATGPT

2 MONTHS

How can we strengthen information integrity?

Mis- and disinformation and hate speech do not exist in a vacuum. They spread when people feel excluded and unheard, when faced with the impacts of economic disparity, and when feeling politically disenchanted. Responses should address these real-world challenges. Efforts to achieve the Sustainable Development Goals are fundamental to building a world in which trust can be restored.

In crafting responses, it is important not to lose sight of the tremendous value digital platforms bring to the world. Platforms have revolutionized real-time mass communication, enabling the spread of life-saving information during natural disasters and pandemics. They have helped to mobilize support for the goals for which the United Nations strives, often proving to be positive forces for inclusion and participation in public life. They have connected geographically disparate communities of people otherwise excluded, including those suffering from rare health conditions, and a vast array of activists working to make the world a better place.

REGULATORY RESPONSES

The question of whether digital platforms can and should be held legally liable for the content that they host has been the subject of lengthy debate. In some contexts, existing laws based on defamation, cyberbullying and harassment have been used effectively to counter threats to information integrity without imposing new restrictions on freedom of expression ([A/77/287](#), para. 44).

In addition, some recent legislative efforts have been made to address the issue at the regional and national levels. These include the framework adopted by the European Union in 2022, comprising the Digital Services Act, the initiative on transparency and targeting of political advertising

and the Code of Practice on Disinformation. The Digital Services Act establishes new rules for users, digital platforms and businesses operating online within the European Union. Measures take aim at illegal online content, goods and services and provide a mechanism for users both to flag illegal content and to challenge moderation decisions that go against them. They require digital platforms to improve transparency, especially on the use and nature of recommendation algorithms, and for larger platforms to provide researchers with access to data.

The Code of Practice on Disinformation sets out principles and commitments for online platforms and the advertising sector to counter the spread of disinformation online in the European Union, which its signatories agreed to implement.⁴¹ These include voluntary commitments to help demonetize disinformation, both by preventing the dissemination of advertisements containing disinformation and by avoiding the placement of advertisements alongside content containing disinformation. Signatories also agreed to label political advertising more clearly, together with details of the sponsor, advertising spend and display period, and to create searchable databases of political advertisements. Furthermore, they committed to share information about malicious manipulative behaviours used to spread disinformation (such as fake accounts, bot-driven amplification, impersonation and malicious deep-fakes) detected on their platforms and regularly update and implement policies to tackle them. Other commitments are focused on empowering users to recognize, understand and flag disinformation, and on strengthening collaboration with fact-checkers and providing researchers with better access to data. The real test of these new mechanisms will be in their implementation.

One of the key aims of the Code of Practice on Disinformation is to improve platform transparency. In February 2023, signatories to the Code of Practice published their first baseline reports on how they are implementing the commitments. The reports provided insight into the extent to which advertising revenue was prevented from flowing to disinformation actors and other detected manipulative behaviour, including a large-scale coordinated effort to manipulate public opinion about the war in Ukraine in several European countries.⁴²

DIGITAL PLATFORM RESPONSES

Digital platforms are very diverse in terms of size, function and structure and have pursued a wide range of responses to tackle harm. Several of the larger platforms have publicly committed to uphold the Guiding Principles on Business and Human Rights,⁴³ but gaps persist in policy, transparency and implementation. Some platforms do not enforce their own standards and, to varying degrees, allow and amplify lies and hate.⁴⁴ Algorithms created to further the platforms' profit-making model are designed to deliberately maximize engagement and monopolize attention, tending to push users towards polarizing or provocative content.

Most digital platforms have some kind of system of self-regulation, moderation or oversight mechanisms in place, yet transparency around content removal policy and practice remains a challenge.⁴⁵ Investment in these mechanisms across regions and languages is extremely patchy and largely concentrated in the global North, as is platforms' enforcement of their own rules. Translation of moderation tools and oversight mechanisms into local languages is incomplete across platforms, a recent survey found.⁴⁶ At the same time, moderation is often outsourced and woefully under-resourced in languages other than English.⁴⁷ Testimony from moderators has raised troubling questions related to mistreatment, labour standards and secondary trauma.⁴⁸ Moderators report being constantly exposed to violent and disturbing content, and being given a matter of seconds to determine if a reported post violates company

policy. Automated content moderation systems can play an essential role, but are exposed to possible bias based on the data and structures used to train them. They also have high rates of error in English and even worse success rates across other languages. A number of digital platforms employ trust and safety, human rights and information integrity teams, yet these experts are often not included at the earliest stages of product development and are often the first jobs cut during cost-saving measures.

DATA ACCESS

Data access for researchers is also an urgent priority on a global scale. Existing research and resources remain heavily skewed towards the United States of America and Europe. With notable exceptions, including the reports on United Nations peacekeeping in Africa and the independent international fact-finding mission on Myanmar,⁴⁹ and some investigative reporting and coverage by journalists,⁵⁰ limited research has been published about the impact on the rest of the world. This is partly because researchers lack access to platforms and their data. The tools needed for effective research of the limited data provided by the platforms also tend to be designed with marketing in mind and are largely prohibitively expensive. A shift by the platforms from an "access by request" approach to "disclosure by default", with necessary safeguards for privacy, would allow researchers to properly evaluate harms.

USER EMPOWERMENT

Civil society groups and academics have conducted extensive research on how best to tackle mis- and disinformation and hate speech while protecting freedom of expression. A number have stressed the need for bottom-up solutions that empower Internet users to limit the impact of online harms on their own communities and decentralize power from the hands of the platforms.

Platform users, including marginalized groups, should be encouraged, included and involved in

the policy space. Youth in particular have a wealth and depth of expertise. As digital natives, young people, in particular young women, and children are already often the targets of mis- and disinformation and hate speech and will be directly affected by emerging and new platforms. Younger users can speak from experience about the differentiated impact of various proposals and their potential flaws. They have also actively contributed to online advocacy and fact-checking efforts.⁵¹

Improved critical thinking skills can make users more resilient against digital manipulation. Specifically, digital literacy teaches users to better evaluate the information that they encounter online and pass it on in a responsible way. A range of United Nations entities have had valuable experiences in this field. The United Nations Verified initiative⁵² has successfully deployed a range of tactics, including targeting messaging for users, pre-bunking – warning users about falsehoods before they encounter them – and digital literacy drives.

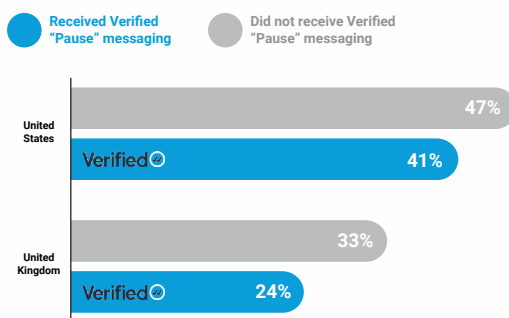
DISINCENTIVES

The current business models of most digital platforms prioritize engagement above human rights, privacy and safety. This includes the monetization of personal data for profit, despite growing evidence of societal harm caused by this business model.

FIGURE III

UNITED NATIONS CAMPAIGNS EFFECTIVE IN COUNTERING MIS- AND DISINFORMATION

Likelihood of sharing fake news (2021)^a



^aBased on research conducted by the Massachusetts Institute of Technology in March 2021

Some civil society groups and researchers have explored avenues for demonetizing and therefore disincentivizing the creation and spread of online mis- and disinformation and hate speech, noting that while freedom of expression is a fundamental human right, profiting from it is not.⁵³ Proposals seek to address the profitability of disinformation, ensure full transparency around monetization of content and independent risk assessments, and disincentivize those involved in online advertising from enabling disinformation.

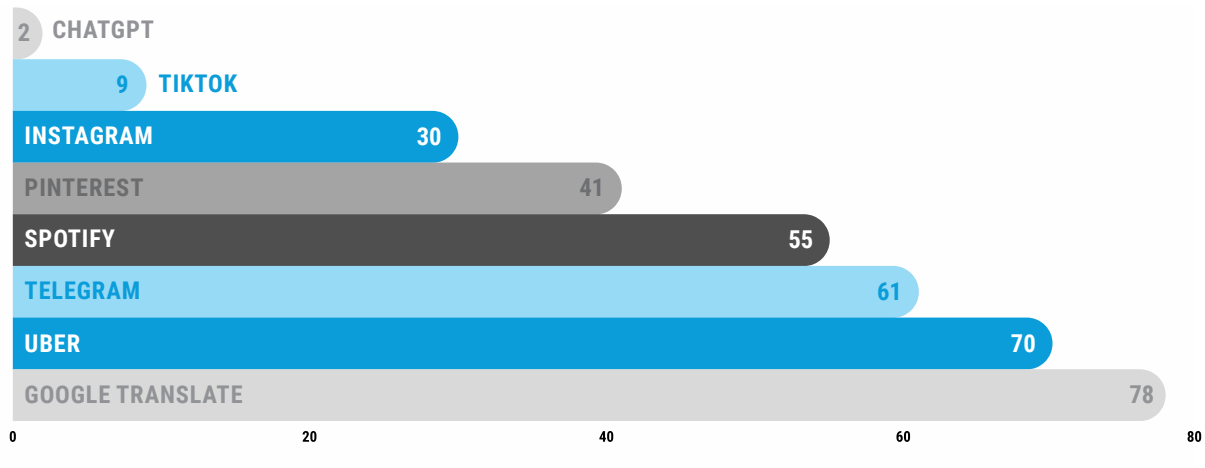
Brands that advertise alongside mis- and disinformation and hate speech risk undermining the effectiveness of their campaigns and, ultimately, their reputations. Advertisers can develop clear policies to avoid inadvertently funding and legitimizing mis- and disinformation and hate speech and help make them unprofitable. Implementation measures can include managing up-to-date inclusion and exclusion lists and using advertisement verification tools. Advertisers can also pressure digital platforms to step up action to protect information integrity and can refrain from advertising with media outlets that fuel hatred and spread disinformation.⁵⁴

INDEPENDENT MEDIA

New measures in dozens of countries continue to undermine press freedom. According to the UNESCO global report for 2022 from its flagship World Trends in Freedom of Expression and Media Development series, 85 per cent of the world's population experienced a decline in press freedom in their country during the preceding five years.⁵⁵ With 2.7 billion people still offline,⁵⁶ a further priority is to strengthen independent media, boost the prevalence of fact-checking initiatives and underpin reliable and accurate reporting in the public interest. Real public debate relies on the facts, told clearly, and reported ethically and independently. Ethical reporters, with quality training and working conditions, have the skills to restore balance in the face of mis- and disinformation. They can offer a vital service: accurate, objective and reliable information about the issues that matter.

FIGURE IV

MONTHS TAKEN TO ADD 100 MILLION MONTHLY ACTIVE USERS FOR CHATGPT AS COMPARED WITH OTHER POPULAR APPS



Source: Similarweb, using data from Sensor Tower

FUTURE-PROOFING

Even as we seek solutions to protect information integrity in the current landscape, we must ensure that recommendations are future-proof, addressing emerging technologies and those yet to come. Launched in November 2022, Open AI’s ChatGPT-3 platform gained 100 million users by January 2023, making it the fastest-growing consumer app in history,⁵⁷ with many other companies racing to develop competing tools. While holding almost unimaginable potential to address global challenges, there are serious and urgent concerns about the equally powerful potential of recent advances in artificial intelligence – including image generators and video deepfakes – to threaten information integrity. Recent reporting and research have shown that generative artificial intelligence tools generated mis- and disinformation and hate speech, convincingly presented to users as fact.⁵⁸

My Envoy on Technology is leading efforts to assess the implications of generative artificial intelligence and other emerging platforms. In doing so we must learn from the mistakes of the past. Digital platforms were launched into the world without sufficient awareness or assessment of the potential damage to societies and individuals. We have the opportunity now to ensure that history

does not repeat itself with emerging technology. The era of Silicon Valley’s “move fast and break things” philosophy must be brought to a close. It is essential that user privacy, security, transparency and safety by design are integrated into all new technologies and products at the outset.

UNITED NATIONS RESPONSES

Steps are also being taken, including by United Nations peace operations and country offices, to monitor, analyse and respond to the threat that mis- and disinformation pose to the delivery of United Nations mandates. The United Nations Strategy and Plan of Action on Hate Speech sets out strategic guidance for the Organization to address hate speech at the national and global levels. In February 2023, UNESCO hosted the Internet for Trust conference to discuss a set of draft global guidelines for regulating digital platforms, due to be finalized later this year.⁵⁹

Together, these initiatives and approaches help to point the way forward for the underlying principles of a United Nations Code of Conduct.

MONTHLY USES OF

#CLIMATESCAM

ON TWITTER

100K

NOVEMBER 2022

COP27 SPIKE

200

FEBRUARY 2022

POSTS PEAK TO

199,300

IN JANUARY 2023

Towards a United Nations Code of Conduct

The United Nations Code of Conduct for Information Integrity on Digital Platforms, which I will put forward, would build upon the following principles:

- Commitment to information integrity
- Respect for human rights
- Support for independent media
- Increased transparency
- User empowerment
- Strengthened research and data access
- Scaled up responses
- Stronger disincentives
- Enhanced trust and safety

These principles have been distilled from the core ideas discussed in the present policy brief, and are in line and interlinked with my policy brief on a Global Digital Compact. Member States will be invited to implement the Code of Conduct at the national level. Consultations will continue with stakeholders to further refine the content of the Code of Conduct, as well as to identify concrete methodologies to operationalize its principles.

The Code of Conduct may draw upon the following recommendations:

Commitment to information integrity

- (a) All stakeholders should refrain from using, supporting or amplifying disinformation and hate speech for any purpose, including to pursue political, military or other strategic goals, incite violence, undermine democratic processes or target civilian populations, vulnerable groups, communities or individuals;

Respect for human rights

- (b) Member States should:

- (i) Ensure that responses to mis- and disinformation and hate speech are consistent with international law, including international human rights law, and are not misused to block any legitimate expression of views or opinion, including through blanket Internet shutdowns or bans on platforms or media outlets;

- (ii) Undertake regulatory measures to protect the fundamental rights of users of digital platforms, including enforcement mechanisms, with full transparency as to the requirements placed on technology companies;

- (c) All stakeholders should comply with the Guiding Principles on Business and Human Rights;

Support for independent media

- (d) Member States should guarantee a free, viable, independent and plural media landscape with strong protections for journalists and independent media, and support the establishment, funding and training of independent fact-checking organizations in local languages;
- (e) News media should ensure accurate and ethical independent reporting supported by quality training and adequate working conditions in line with international labour and human rights norms and standards;

Increased transparency

- (f) Digital platforms should:
 - (i) Ensure meaningful transparency regarding algorithms, data, content moderation and advertising;
 - (ii) Publish and publicize accessible policies on mis- and disinformation and hate speech, and report on the prevalence of coordinated disinformation on their services and the efficacy of policies to counter such operations;
- (g) News media should ensure meaningful transparency of funding sources and advertising policies, and clearly distinguish editorial content from paid advertising, including when publishing to digital platforms;

User empowerment

- (h) Member States should ensure public access to accurate, transparent, and credibly sourced government information, particularly information that serves the public interest, including all aspects of the Sustainable Development Goals;
- (i) Digital platforms should ensure transparent user empowerment and protection, giving people greater choice over the content that they see and how their data is used. They should enable users to prove identity and authenticity free of monetary or privacy trade-offs and establish transparent user complaint and reporting processes supported by independent, well publicized and accessible complaint review mechanisms;
- (j) All stakeholders should invest in robust digital literacy drives to empower users of all ages to better understand how digital platforms work, how their personal data might be used, and to identify and respond to mis- and disinformation and hate speech. Particular attention should be given to ensuring that young people, adolescents and children are fully aware of their rights in online spaces;

Strengthened research and data access

- (k) Member States should invest in and support independent research on the prevalence and impact of mis- and disinformation and hate speech across countries and languages, particularly in underserved contexts and in languages other than English, allowing civil society and academia to operate freely and safely;
- (l) Digital platforms should:

(i) Allow researchers and academics access to data, while respecting user privacy. Researchers should be enabled to gather examples and qualitative data on individuals and groups targeted by mis- and disinformation and hate speech to better understand the scope and nature of harms, while respecting data protection and human rights;

(ii) Ensure the full participation of civil society in efforts to address mis- and disinformation and hate speech;

Scaled-up responses

(m) All stakeholders should:

(i) Allocate resources to address and report on the origins, spread and impact of mis- and disinformation and hate speech, while respecting human rights norms and standards and further invest in fact-checking capabilities across countries and contexts;

(ii) Form broad coalitions on information integrity, bringing together different expertise and approaches to help to bridge the gap between local organizations and technology companies operating at a global scale;

(iii) Promote training and capacity-building to develop understanding of how mis- and disinformation and hate speech manifest and to strengthen prevention and mitigation strategies;

Stronger disincentives

(n) Digital platforms should move away from business models that prioritize engagement above human rights, privacy and safety;

(o) Advertisers and digital platforms should ensure that advertisements are not placed

next to online mis- or disinformation or hate speech, and that advertising containing disinformation is not promoted;

(p) News media should ensure that all paid advertising and advertorial content is clearly marked as such and is free of mis- and disinformation and hate speech;

Enhanced trust and safety

(q) Digital platforms should:

(i) Ensure safety and privacy by design in all products, including through adequate resourcing of in-house trust and safety expertise, alongside consistent application of policies across countries and languages;

(ii) Invest in human and artificial intelligence content moderation systems in all languages used in countries of operation, and ensure content reporting mechanisms are transparent, with an accelerated response rate, especially in conflict settings;

(r) All stakeholders should take urgent and immediate measures to ensure the safe, secure, responsible, ethical and human rights-compliant use of artificial intelligence and address the implications of recent advances in this field for the spread of mis- and disinformation and hate speech.

Next Steps

- The United Nations Secretariat will undertake broad consultations with a range of stakeholders on the development of the United Nations Code of Conduct, including mechanisms for follow-up and implementation. This could include the establishment of an independent observatory made up of recognized experts to assess the measures taken by the actors who commit to the Code of Conduct, and other reporting mechanisms.
- To support and inform the Code, the United Nations Secretariat may carry out in-depth studies to enhance understanding of information integrity globally, especially in underresearched parts of the world.
- The Secretary-General will establish dedicated capacity in the United Nations Secretariat to scale up the response to online mis- and disinformation and hate speech affecting United Nations mandate delivery and substantive priorities. Based on expert monitoring and analysis, such capacity would develop tailored communication strategies to anticipate and/or rapidly address threats before they spiral into online and offline harm, and support capacity-building of United Nations staff and Member States. It would support efforts of Member States, digital platforms and other stakeholders to adhere to and implement the Code when finalized.

Conclusion

Strengthening information integrity on digital platforms is an urgent priority for the international community. From health and gender equality to peace, justice, education and climate action, measures that limit the impact of mis- and dis-information and hate speech will boost efforts to achieve a sustainable future and leave no one behind. Even with action at the national level, these problems can only be fully addressed through stronger global cooperation. The core ideas outlined in this policy brief demonstrate that the path towards stronger information integrity needs to

be human rights-based, multi-stakeholder, and multi-dimensional. They have been distilled into a number of principles to be considered for a United Nations Code of Conduct for Information Integrity on Digital Platforms that would provide a blueprint for bolstering information integrity while vigorously upholding human rights. I look forward to collaborating with Member States and other stakeholders to turn these principles into tangible commitments.

Annex

CONSULTATIONS WITH MEMBER STATES AND OTHER RELEVANT STAKEHOLDERS

The ideas in the present policy brief draw on the proposals outlined in the report entitled “Our Common Agenda” (A/75/982), which benefited from extensive consultations with Member States, the United Nations system, thought leaders, young people and civil society actors from all around the world. The policy brief responds, in particular, to the rich and detailed reflections of Member States and other stakeholders on Our

Common Agenda over the course of 25 General Assembly discussions.

In advance of the publication of the present policy brief, consultations were carried out with Member States, including through an informal briefing to the Committee on Information, which all non-Committee members were invited to join. Discussions were also held with civil society partners, academics, experts and the private sector, including technology companies.

Broad consultations will be conducted in the development of the Code of Conduct, ahead of the Summit of the Future.

Endnotes

- 1 [A/HRC/42/50](#); [A/77/287](#); [A/HRC/51/53](#); United Nations, “Statement by Alice Wairimu Nderitu, Special Adviser on the Prevention of Genocide, condemning the recent escalation of fighting in Ethiopia”, press release, 19 October 2022; Office of the United Nations High Commissioner for Human Rights (OHCHR), “Myanmar: Social media companies must stand up to junta’s online terror campaign say UN experts”, press release, 13 March 2023; OHCHR, “Freedom of speech is not freedom to spread racial hatred on social media: UN experts”, statement, 6 January 2023; Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, “#JournalistsToo: women journalists speak out”, 24 November 2021; and OHCHR, “Sri Lanka: Experts dismayed by regressive steps, call for renewed UN scrutiny and efforts to ensure accountability”, press release, 5 February 2021.
- 2 [A/HRC/47/25](#), para. 15.
- 3 Kalina Bontcheva and Julie Posetti, eds., *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression – Broadband Commission Research Report on “Freedom of Expression and Addressing Disinformation on the Internet”* (Geneva, International Telecommunication Union (ITU); Paris, UNESCO, 2020).
- 4 See United Nations, “Countering disinformation”, available at www.un.org/en/countering-disinformation, and [A/77/287](#).
- 5 Available at www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf.
- 6 The United Nations is currently undertaking a study to examine the interlinkages and relationship between mis- and disinformation and hate speech, and where these related but distinct phenomena converge and diverge at both the conceptual and operational levels.
- 7 The European Commission defines online platforms at “Shaping Europe’s digital future: online platforms”, 7 June 2022. Available at <https://digital-strategy.ec.europa.eu/en/policies/online-platforms>.
- 8 See OHCHR, “Moderating online content: fighting harm or silencing dissent?”, 23 July 2021.
- 9 See United Nations, “Countering disinformation”, and [A/77/287](#).
- 10 [A/HRC/47/25](#).
- 11 Stephanie Kirchgaessner and others, “Revealed: the hacking and disinformation team meddling in elections”, *The Guardian*, 14 February 2023.
- 12 Alexandre Alaphilippe and others, “Doppelganger – media clones serving Russian propaganda”, EU DisinfoLab, 27 September 2022.
- 13 United Nations Economist Network, “New economics for sustainable development: attention economy”.
- 14 Twitter, “About Twitter Blue”; and Meta, “Testing Meta Verified to help creators establish their presence”, 17 March 2023.
- 15 [A/77/287](#).
- 16 As at February 2023, 173 Member States were States parties to the International Covenant on Civil and Political Rights.
- 17 Limitations on freedom of expression must meet the following well-established conditions: legality, that is, restrictions must be provided by law in a manner that distinguishes between lawful and unlawful expression with sufficient precision; necessity and proportionality, that is, the limitation demonstrably imposes the least burden on the exercise of the right and actually protects, or is likely to protect, the legitimate State interest at issue; and legitimacy, that is, to be lawful, restrictions must protect only those interests enumerated in article 19 (3) of the International Covenant on Civil and Political Rights
- 18 These include audiovisual communications in Côte d’Ivoire, Morocco and Tunisia, and monitoring of incitement to violence by the United Nations Multidimensional Integrated Stabilization Mission in the Central African Republic.
- 19 Aleksu Knuutila, Lisa-Maria Neudert and Philip N. Howard, “Who is afraid of fake news? Modeling risk perceptions of misinformation in 142 countries”, *Harvard Kennedy School (HKS) Misinformation Review*, vol. 3, No. 3 (April 2022).
- 20 ITU, *Measuring the Information Society* (Geneva, 2013).
- 21 United Nations Children’s Fund (UNICEF), “Protecting children online”, 23 June 2022. Available at www.unicef.org/protection/violence-against-children-online.
- 22 UNESCO, working papers on digital governance and the challenges for trust and safety. Available at www.unesco.org/en/internet-conference/working-papers.
- 23 See Julie Posetti and Kalina Bontcheva, “Disinfodemic: deciphering COVID-19 disinformation” policy brief 1, (Paris, UNESCO, 2020), and “Disinfodemic: dissecting responses to COVID-19 disinformation” policy brief 2, (Paris, UNESCO, 2020).
- 24 Center for Countering Digital Hate, *Pandemic Profiteers: The Business of Anti-Vaxx* (2021).
- 25 Michael A Gisondi and others, “A deadly infodemic: social media and the power of COVID-19 misinformation”, *Journal of Medical Internet Research*, vol. 24, No. 2 (February 2022).
- 26 [A/77/288](#).
- 27 [A/77/287](#), para. 6.
- 28 In 2018, an independent international fact-finding mission appointed by the Human Rights Council declared Facebook to be “the leading platform for hate speech in Myanmar” ([A/HRC/42/50](#), para. 72).

- 29 See United Nations News, “[Hate speech: a growing, international threat](#)”, 28 January 2023, and “[Digital technology, social media fuelling hate speech like never before, warns UN expert](#)”, 20 October 2022.
- 30 See John Cook, “Deconstructing climate science denial”, in *Research Handbook in Communicating Climate Change*, David C. Holmes and Lucy M. Richardson, eds. (Cheltenham, United Kingdom, Edward Elgar, 2020). Cook reported that Abraham et al. (2014) had summarized how papers containing denialist claims, such as claims of cooling in satellite measurements or estimates of low climate sensitivity, have been robustly refuted in the scientific literature. Similarly, Benestad et al. (2016) attempted to replicate findings in contrarian papers and found a number of flaws such as inappropriate statistical methods, false dichotomies, and conclusions based on misconceived physics.
- 31 Global Witness, “[The climate divide: how Facebook’s algorithm amplifies climate disinformation](#)”, 28 March 2022.
- 32 Analysis by the Department of Global Communications, using data from Talkwalker.
- 33 Jeffrey A. Hicke and others, “North America”, in Intergovernmental Panel on Climate Change, *Climate Change 2022: Impacts, Adaptation and Vulnerability*, Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (Cambridge, United Kingdom, Cambridge University Press, 2022).
- 34 Mei Li, Gregory Trencher and Jusen Asuka, “[The clean energy claims of BP, Chevron, ExxonMobil and Shell: a mismatch between discourse, actions and investments](#)”, PLOS ONE, issue 17, No. 2 (February 2022).
- 35 Robert J. Brulle and Carter Werthman, “[The role of public relations firms in climate change politics](#)”, *Climatic Change*, vol. 169, No. 1–2 (November 2021). According to the Global Disinformation Index, a non-profit watchdog, tech industry advertisers provided \$36.7 million to 98 websites carrying climate disinformation in English in 2021. A November 2022 report from the Center for Countering Digital Hate, a campaign group, revealed that, on Google alone, nearly half of the \$23.7 million spent on search ads by oil and gas companies in the last two years targeted search terms on environmental sustainability. Research from InfluenceMap found 25,147 misleading ads from 25 oil and gas sector organizations on Facebook’s platforms in the United States of America in 2020, with a total spend of \$9,597,376. So far the response has been incommensurate with the scale of the problem.
- 36 See General Assembly resolution [76/227](#); Human Rights Council resolution [49/21](#); and European Union External Action, “[Tackling disinformation, foreign information manipulation and interference](#)”, 27 October 2021.
- 37 Lucina Di Meco, “[Monetizing misogyny: gendered disinformation and the undermining of women’s rights and democracy globally](#)”, #ShePersisted, February 2023.
- 38 See Andrew Puddephatt, “[Social media and elections](#)”, Cuadernos de Discusión de Comunicación e Información, No. 14 (Montevideo, UNESCO, 2019); and Julie Posetti and others, “[The chilling: global trends in online violence against women journalists](#)”, research discussion paper (UNESCO, 2021).
- 39 EU Disinfo Lab, “[The role of “media” in producing and spreading disinformation campaigns](#)”, 13 October 2021.
- 40 See United Nations News, “[Social media poses ‘existential threat’ to traditional, trustworthy news: UNESCO](#)”, 10 March 2022; and Anya Schiffrin and others, “[Finding the funds for journalism to thrive: policy options to support media viability](#)”, World Trends in Freedom of Expression and Media Development (Paris, UNESCO, 2022).
- 41 European Commission, “[Shaping Europe’s digital future: the 2022 Code of Practice on Disinformation](#)”, 4 July 2022.
- 42 See the remarks by the European Commission Vice-President for Values and Transparency, Věra Jourová, in European Commission, “[Code of Practice on Disinformation: new Transparency Centre provides insights and data on online disinformation for the first time](#)”, daily news, 9 February 2023. Available at https://ec.europa.eu/commission/presscorner/detail/en/mex_23_723.
- 43 Available at <https://unglobalcompact.org/library/2>.
- 44 Center for Countering Digital Hate and Human Rights Campaign, “[Digital hate: social media’s role in amplifying dangerous lies about LGBTQ+ people](#)”, 10 August 2022.
- 45 See Andrew Puddephatt, “[Letting the sun shine in: transparency and accountability in the digital age](#)”, World Trends in Freedom of Expression and Media Development (Paris, UNESCO, 2021).
- 46 Whose Knowledge?, Oxford Internet Institute and The Centre for Internet and Society, *State of the Internet’s Languages Report (2022)*.
- 47 [A/HRC/38/35](#).
- 48 Billy Perrigo, “[Inside Facebook’s African sweatshop](#)”, *Time*, 17 February 2022.
- 49 [A/HRC/42/50](#).
- 50 Notable examples include Maria Ressa, *How to Stand up to a Dictator* (New York, HarperCollins, 2022); and Max Fischer, *The Chaos Machine* (New York, Little, Brown and Company, 2022).
- 51 See UNICEF, “[Young reporters fact-checking COVID-19 information](#)”.
- 52 See <https://shareverified.com/>.
- 53 The Global Disinformation Index, a non-profit group, tracks advertising placed together with disinformation. The United Nations has been a victim of this practice, with the Global Disinformation Index having found UNICEF advertisements placed alongside anti-vaccine articles, and Office of the United Nations High Commissioner for Refugees advertisements together with anti-refugee content.
- 54 Conscious Advertising Network, manifestos. Available at www.consciousadnetwork.com/the-manifestos/.
- 55 UNESCO, *Journalism is a Public Good: World Trends in Freedom of Expression and Media Development – Global Report 2021/2022* (Paris, 2022).
- 56 ITU, “[Facts and figures 2021: 2.9 billion people still offline](#)”, 29 November 2021. The [global digital compact](#) to be taken up by Member States at the Summit of the Future, to be held in 2024, will outline shared principles for an open, free and secure digital future for all (see www.un.org/techenvoy/global-digital-compact).
- 57 Krystal Hu, “[ChatGPT sets record for fastest-growing user base – analyst note](#)”, Reuters, 2 February 2023.
- 58 See Center for Countering Digital Hate, “[Misinformation on Bard, Google’s new AI chat](#)”, 5 April 2023; and Tiffany Hsu and Stuart A. Thompson, “[Disinformation researchers raise alarms about A.I. chatbots](#)”, *The New York Times*, 13 February 2023.
- 59 The draft guidelines are available at www.unesco.org/en/internet-conference.

