Organization for Security and Co-operation in Europe
Office of the Representative on Freedom of the Media

Artificial Intelligence and Disinformation as a Multilateral Policy Challenge

Brief Paper for the Expert Meeting organized by the Office of the OSCE Representative on Freedom of the Media on 7 December 2021

Prepared by Deniz Wagner
Adviser to the OSCE Representative on Freedom of the Media

Vienna, November, 2021

Table of Contents

Introduction

Systemic issues arguably do not arise from disinformation alone; it is the dissemination through artificial intelligence (AI) that plays a key role in amplifying it from individual content to a scale that produces and exacerbates systemic consequences endangering peace and security.

As has been recognized in the first brief paper on *international law and policy on disinformation in the context of freedom of the media*, the international problem of how to counteract the dissemination of false reports and information detrimental to peace, security and cooperation has existed for a hundred years. There is a body of international law that addresses disinformation, especially in the context of the harm it has on international relations. Today, the desire to find a solution has risen in line with the growth of the media's influence, intensified by the role that social media plays in informing the public.[1]

AI plays a central role for online platforms. It is becoming, if not already, a key instrument for shaping and arbitrating online information spaces. Through the design and deployment of AI, online platforms are also directly influencing people's opinions and expression, which, at scale, is also leading to systemic and structural challenges for comprehensive security. A particular challenge is the reckless and pervasive spread of disinformation, with AI acting as its amplifier.

At the core of this challenge is information saturation leading to a need for the structuring and prioritizing of information that is no longer possible manually. The sheer volume of content, the overwhelming number of narratives and counternarratives, and the pace of the news cycle are difficult to rationalize, digest and meaningfully interpret without technical assistance. What we are witnessing is a new form of attack on freedom of expression. While censorship focuses on the suppression of speech, new tactics do quite the opposite and flood the online space with all kinds of speech, including multitudes of false, inaccurate and misleading information. This *weapon of mass distraction*[2] is proving incredibly effective in creating chaos and distrust in institutions.

Artificial Intelligence

In 1842, Ada Lovelace had declared that "the Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis, but it has no power of anticipating any analytical relations or truths."[3]

In many ways, this description holds true even today. Most international descriptions refer to AI in one way or another as data-driven machine-based systems, operating with varying levels

---

[1] Rikhter, OSCE RFOM Brief Paper on International Law and Policy on Disinformation in the Context of Freedom of the Media (2021); https://www.osce.org/files/f/documents/8/a/485606.pdf

[2] Christina Nemr and William Gangware, Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age (2019); https://www.state.gov/wp-content/uploads/2019/05/Weapons-of-Mass-Distraction-Foreign-State-Sponsored-Disinformation-in-the-Digital-Age.pdf

[3] BBC News, A Point of View: Will machines ever be able to think? (2013); https://www.bbc.com/news/magazine-24565995

of autonomy, to make predictions, recommendations, or decisions for a given set of human defined objectives, with the aim to ultimately influence offline or virtual environments.[4]

AI is therefore arguably not "machine intelligence" but human intelligence embedded into a data structure, held and processed by a machine to expose information or perform processes at speed and scale, like a Rube Goldberg Machine[5] - but instead of performing a simple task in a complicated manner, it aims to perform a complicated task in a simplified manner.

The term "artificial intelligence" can encompass a variety of different concepts of automated processes. A particular component is the "algorithm" – a sequence of commands in the form of computer code, that carries out a set of instructions, generating output from a given input in a clearly defined format.[6]

AI systems are frequently used for the large-scale processing of user data and profiling, which pose risks to the rights to privacy and freedom of expression.

Ultimately, AI is used to support the dissemination of online content to audiences (content curation) as well as to filter content for identifying and removing or deprioritizing illegal or otherwise undesirable online content (content moderation). These processes influence the information basis of how society interacts online today.

Content Curation – Amplifying Disinformation

The internet hosts an immeasurable amount of content. More than 500 hours of video is uploaded to YouTube every minute, almost 9 million photos uploaded to Facebook every hour, and more than 500 million stories are posted on Instagram in a day.[7]

Audiences need help to navigate this abundance of online content. Content recommender systems can be useful in that regard – these AI-led systems search through endless content to provide personalized selections that are predicted as relevant for the user.[8] Recommender systems play a central role across the most popular websites and platforms on the internet.[9] However, this is by no means an impartial process. The operation of recommender systems is driven by the underlying design choices and business interest of many online social media platforms and search engines, by exploiting data about user behaviour to manipulate their

[4] Bukovska, OSCE RFOM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (2019); https://www.osce.org/files/f/documents/9/f/456319_0.pdf; Krivokapic, OSCE Non-Paper on the Impact of Artificial Intelligence on Freedom of Expression (2019); https://www.osce.org/representative-on-freedom-of-media/447829; and OECD Recommendation on Artificial Intelligence (2020); https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#backgroundInformation

[5] A Rube Goldberg machine is a contraption that uses a chain reaction to carry out a simple task; It performs a very basic job in a complicated way.

[6] Bukovska, OSCE RFOM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (2019); https://www.osce.org/files/f/documents/9/f/456319_0.pdf

[7] Domo, Data Never Sleeps 8.0, (2020); https://www.domo.com/learn/infographic/data-never-sleeps-8

[8] Llansó, Hoboken, Leerssen, Harambam, Artificial Intelligence, Content Moderation, and Freedom of Expression (2020); https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

[9] Cobbe and Singh, Regulating Recommending: Motivations, Consideration, and Principles (2019); https://ejlt.org/index.php/ejlt/article/view/686/979, Table 1

attention to ultimately increase advertising revenue.[10] These systems of targeted advertising are therefore frequently programmed to prioritize click-worthy content, rather than newsworthy content.[11]

Fundamentally, recommender systems are typically used by platforms to show users whatever the algorithms predict will drive engagement, revenue, and market position,[12] often with little to no regard for what the material being disseminated actually is[13]. This can financially incentivise the creation and promotion of content that is tabloid, controversial, or otherwise produces an emotional response, including misinformation and disinformation.

A number of studies have found that falsehoods are spread online significantly faster, deeper, and more broadly than the truth, in all categories of information. One MIT study discovered that stories based on false or misleading information are 70 percent more likely to be retweeted than true stories; and that it takes stories based on accurate information about six times as long to reach 1,500 people as it does for false information to reach the same number of people.[14]

Several years ago, sociologist Zeynep Tufekci set out a thesis that YouTube is a so-called "radicalization engine"[15]. Tufekci argues that Google's algorithms seem to be wired by the assumption that people are drawn to increasingly more extreme content than the first piece of content they originally viewed on the platform. This, she posits, is a "computational exploitation of a natural human desire: to look 'behind the curtain,' to dig deeper into something that engages us. As we click and click, we are carried along by the exciting sensation of uncovering more secrets and deeper truths. YouTube leads viewers down a rabbit hole of extremism, while Google racks up the ad sales."[16]

YouTube is the second most visited website in the world, with 70 percent of its user engagement resulting from recommended videos (in other words, from viewing videos that were not specifically sought after in the search function but rather through clicking on a recommended video or continuing to watch the "up next" videos that appear once a video is done). So, the scale, regarding the amount of disinformation spread on this platform can be immense.

A recent study of online video content, and YouTube in particular, discovered that people in non-English speaking countries are most exposed to content deemed disturbing or harmful.[17] Meanwhile, Facebook's AI-powered content moderation cannot read many languages of

---

[10] Ricci, Rokach, Shapira, Recommender Systems Handbook (2015); Springer
[11] Bukovska, OSCE RFOM Policy Paper on Freedom of the Media and Artificial Intelligence (2020); https://www.osce.org/files/f/documents/4/5/472488.pdf
[12] Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (2019); Profile Books
[13] Cobbe and Singh, Regulating Recommending: Motivations, Consideration, and Principles (2019); https://ejlt.org/index.php/ejlt/article/view/686/979
[14] Dizikes, Study: On Twitter, false news travels faster than true stories, MIT News (2018); https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308
[15] Tufekci, YouTube, the Great Radicalizer (2018); https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html
[16] Friedersdorf, YouTube Extremism and the Long Tail, The Atlantic (2018); https://www.theatlantic.com/politics/archive/2018/03/youtube-extremism-and-the-long-tail/555350/
[17] Mozilla, YouTube Regrets: A crowdsourced investigations into YouTube's recommendation algorithm (2021); https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf

countries and regions the platform operates in.[18] Such blind spots make online platforms particularly susceptible to letting bad actors post harmful content, including disinformation.[19]

Another substantial aspect of this problem are the algorithms behind these recommender systems; Research confirms the ability of algorithmic content curation to have a significant influence[20] on societies. With recommender systems acting as a facilitator for the spread of disinformation in the digital sphere, its influence and impact on societies in palpable.

The use of AI tools to target and manipulate people, having detrimental implications for elections, democracy and social cohesion has been elaborated through a number of revelations, including the Cambridge Analytica Scandal of 2018 and most recently the Facebook Files of Frances Haugen.[21]

Polarisation

Based on the above assessment, AI targets individuals with purposely tailored content. Such made-to-measure content raises concerns over fragmentation of information spaces and polarisation in various ways including by narrowing the choice of content visible to users leading to partial information blindness (so-called 'filter bubbles') or by recommending content that reinforces the users' existing views (so-called 'echo chambers'). While it seems comforting to solely engage with content that bolsters existing views, it can ultimately lead to perpetual tribalism[22], and severely skew exposure to the surrounding world.

Of course, some degree of polarisation is inescapable, and can hardly be blamed solely on AI. Many forms of legacy media have historically also catered to an audience of like-minded i.e. television channels or newspapers with clear political alignments. Nevertheless, audiovisual media is highly regulated through robust laws and independent regulators acting as watchdogs ensuring news programs report factually. Print press is also bound by self-regulatory professional codes of ethics. Ultimately, in most cases, legacy media do not act with a carte blanche to concoct lies and spread disinformation.

Online platforms, however, are not bound by journalistic ethics; and so factual accuracy cannot be taken for granted. This is a crucial acknowledgement for the growing number of people relying primarily (if not solely) on online platforms for news consumption.[23]

---

[18] Canales, Facebook's AI moderation reportedly can't interpret many languages, leaving users in some countries more susceptible to harmful posts, INSIDER (2021); https://www.businessinsider.com/facebook-content-moderation-ai-cant-speak-all-languages-2021-9

[19] The Report of the independent international fact-finding mission on Myanmar presented at the 39th session of the UN Human Rights Council in 2018 para. 74 touches upon the significant role of social media, and regrets that Facebook is unable to provide country-specific data about the spread of hate speech on its platform, which is imperative to assess the adequacy of its response. See https://undocs.org/en/A/HRC/39/64

[20] For example, Facebook published research in 2014 which shows that it can actively influence users' emotional state by tweaking its algorithm. See https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds

[21] The Wall Street Journal, The Facebook Files (2021); https://www.wsj.com/articles/the-facebook-files-11631713039 and The Guardian, The Cambridge Analytica Files (2018); https://www.theguardian.com/news/series/cambridge-analytica-files

[22] https://www.theguardian.com/science/blog/2017/dec/04/echo-chambers-are-dangerous-we-must-try-to-break-free-of-our-online-bubbles

[23] An estimated 61% of millennials garner news primarily through social media. See https://www.pewresearch.org/fact-tank/2015/06/01/political-news-habits-by-generation/

The technological curation of our information space by AI fundamentally affects the way we encounter ideas and information online, risking to impede information pluralism.

Inauthentic behaviour

The European Commission states that a coordinated use of fake accounts or other forms of inauthentic behaviour to artificially boost content online is a clear indicator of the intention to use false or misleading information to cause harm.[24]

Earlier this year, Facebook released a threat report on the state of influence operations on its platform between 2017-2020.[25] The report defines influence operations as "coordinated efforts to manipulate or corrupt public debate for a strategic goal." In its report, Facebook says it has uncovered evidence of more than 150 coordinated inauthentic behavior campaigns in more than 50 countries since 2017.

AI tools used for inauthentic behaviour (such as bot armies) help amplify the spread and impact of disinformation; and recent years have seen a surge of bots and trolls aiming to manipulate public discourse on critical issues like elections or the coronavirus pandemic.

Referred to as 'weapons of mass distraction', 'filling the zone' or 'opening the floodgates', AI-powered bots are used to overwhelm the online information to drown out the visibility of public interest content.

This is eroding genuine public debate, fueling disengagement and distrust in democratic institutions.

Moreover, AI tools can also be used by malicious actors to target, and attempt to silence specific dissident voices online. Examples of this are coordinated harassment campaigns against journalists that deceptively resemble grassroots movements, with AI-driven distribution systems enhancing the virality of such online attacks.[26]

Combatting disinformation through content moderation

Online platforms are increasingly called upon to play a more active role in helping governments grapple with disinformation. Over the last few years, several countries have put increasing pressure on online platforms to automate content moderation processes,[27] particularly by pushing them to remove content within a very short time period. Germany's Network Enforcement Act (Netzwerkdurchsetzungsgesetz or - NetzDG)[28] adopted in 2017 obliges

[24] European Commission, Tackling COVID-19 disinformation - Getting the facts right (2020); https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020JC0008
[25] Facebook, Threat Report: The State of Influence Operations 2017-2020 (2021); https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf
[26] Haas, OSCE RFOM Policy paper on freedom of the media and artificial intelligence (2020); https://www.osce.org/files/f/documents/4/5/472488.pdf
[27] See concise list of AI regulatory initiatives: European Agency for Fundamental Rights, AI policy initiatives (2016-2019), available at: https://fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights/ai-policy-initiatives
[28] The OSCE Representative on Freedom of the Media reviewed this legislation and warned that warned that it could have a disproportionate effect on freedom of expression. See https://www.osce.org/fom/347651

content that is "clearly illegal" to be removed within 24 hours after receiving a user complaint[29], and the 2016 EU Code of Conduct on countering illegal hate speech online calls to remove or disable access to illegal hate speech in less than 24 hours[30].

There can never be enough human moderators to fully grasp the entirety of the content that needs to be moderated online, not to mention the toll such moderation takes on individuals.[31] AI therefore becomes a necessity, at the very least, to facilitate content moderation.

AI tools are deployed by online platforms to police content across an array of issues, including misinformation and inauthentic behaviour, at scale. Algorithmic content moderation techniques are used to detect potentially problematic content, to evaluate and enforce a decision to tag, label and flag, demonetize, demote, or prioritize certain content, based on legality and/or potential harms. Text analysis and image analysis are the two most commonly used content moderation techniques for combatting potentially illegal or harmful content online. However, the capabilities of such automated content analysis are limited, and challenges linked to its use are manifold.

First, there is a common and incorrect assumption that technology is neutral. Vulnerabilities can start from the outset, with the design of a machine-learning algorithm. A machine-learning model is developed based on a set of training data, decided by humans; it learns, replicates and builds upon what humans instill on it to learn. As such, AI design ingrains choices made by its creators. Humans are prone to bias; they are often innate and an inescapable part of human nature, influenced by our environment and experiences[32]. Due to the innate nature, biases of humans involved in developing AI systems, and those embedded within provided data, will likely be reproduced throughout the lifecycle of the AI system, and amplified through them, at scale.

This phenomenon commonly known as 'algorithmic bias' embeds racial, gendered, class-based and regional discrimination. Joy Buolamwini has coined the term "the coded gaze" referring to algorithmic bias as the "embedded views that are propagated by those who have the power to code systems"[33]. It is also particularly difficult to measure, especially since AI technology operates in a corporate black box. Typically, it is unknown how a particular AI system was designed, what data it was trained on, or how it functions. It is only through the outcome effect of an AI system that we know of many examples, particularly of racial and gendered bias. Such examples include Google's photo recognition software, which labeled photos of Black people as "gorillas";[34] Flickr's auto-tagging system, which labeled concentration camp locations on a

---

[29] See https://www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html

[30] See https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en#theeucodeofconduct

[31] Online platforms hire (and outsource) thousands of human moderators do review potentially harmful and illegal content. This work has been exposed as psychologically scarring, and often under precarious labor arrangements. This human cost provides a strong argument *for* automation. 'The Cleaners' a documentary by Block and Riesewick gives a powerful depiction of these harms. See also: Chotiner, The Underworld of Online Content Moderation, The NewYorker (2019); https://www.newyorker.com/news/q-and-a/the-underworld-of-online-content-moderation

[32] Lebowitz and Lee, 20 cognitive biases that screw up your decisions, INSIDER (2015); https://www.businessinsider.com/cognitive-biases-that-affect-decisions-2015-8

[33] Buolamwini, Fighting the "coded gaze", Ford Foundation (2018); https://www.fordfoundation.org/just-matters/just-matters/posts/fighting-the-coded-gaze/

[34] Vinent, Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech, THE VERGE (2018); https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

map as "sport" or "jungle gym";[35] and Nikon's camera software mislabels images of east-Asian people as blinking.[36]

But to make matters more complicated, we do not even know in many instances whether AI was used in the first place. This lack of transparency around the development and deployment of AI makes it particularly challenging to understand, and address, algorithmic bias.

Linked to this challenge is a lack of diversity in the creators of AI systems. A study of 177 Silicon Valley tech companies showed that ten large technology companies in Silicon Valley did not employ a single black woman in 2018, three had no black employees at all, and six did not have a single female executive[37]. AI tools curating the online information space and moderating it are developed, or commissioned, by these companies and its predominantly male, white, and able-bodied teams, and this under-/misrepresentation of minority or marginalized communities in the development of AI systems means that the AI system developed does not cater to their needs or the needs of their communities. In that regard, it is ineffective to focus on the technology alone as it negates the political and societal systems in which it is developed and operates.[38]

AI is unable to effectively understand or interpret context, nor is it able to understand the intention of the user who posted said content, or in some circumstances, even linguistic, sociological or political context of the post in question. So without a guarantee of human review, such content moderation will almost certainly result in illegitimate restrictions.

Furthermore, AI tools may not be able to ascertain whether certain content is genuinely illegal or harmful. Whether specific content amounts to a violation of an online platform's terms of service (or in some cases, a violation of law) depends on context that the AI technology is unable to incorporate in its evaluation.

Moreover, AI technology is only as good as the datasets used for its training. If these datasets do not include examples of speech in different languages, different communities, and in particular marginalized voices, the resulting technology will fail in its ability to moderate these groups' content, and risks reinforcing as well as deepening existing bias against underrepresented communities.

With this in mind, AI tools will always have so-called 'false positives', whereby content is wrongly classified as objectionable, and 'false negatives' whereby content that should have been classified as objectionable falls through the gaps. Consequently, for many, moderation can be an injustice.

From a freedom of expression perspective, such 'false positives' lead to censorship of legitimate content, while 'false negatives' result in a failure to address harms of disinformation,

---

[35] Dewey, Google Maps' White House glitch, Flickr auto-tag, and the case of the racist algorithm, The Washington Post (2015); https://www.washingtonpost.com/news/the-intersect/wp/2015/05/20/google-maps-white-house-glitch-flickr-auto-tag-and-the-case-of-the-racist-algorithm/

[36] Rose, Are Face-Detection Cameras Racist?, TIME (2010); http://content.time.com/time/business/article/0,8599,1954643,00.html

[37] Rangarajan, Here's the clearest picture of Silicon Valley's diversity yet: It's bad. But some companies are doing less bad, Reveal News (2018); Bay Area tech diversity: White men dominate Silicon Valley (revealnews.org)

[38] Digital Freedom Fund, Decolonising Digital Rights; Decolonising Digital Rights: Why It Matters and Where Do We Start? – Digital Freedom Fund

thereby creating a chilling effect on the ability of targeted individuals or communities' ability to engage online.[39]

Once AI systems are deployed, there is also a lack of adequate oversight and of due process. Without appropriate mechanisms for complaint, review, and appeal, the actions taken as a result of algorithmic decision making may violate the right to freedom of expression.[40]

With concern about the power, scale and impact of AI systems, which give rise to a cluster of concerns, particularly regarding their role in the widescale dissemination of disinformation, there is a need to secure 'algorithmic accountability'. Accountability is vital for establishing avenues of redress, and thereby, protect human rights and dignity.[41]

Conceptual challenge of disinformation

A key challenge in addressing disinformation is the impossibility of always drawing a clear line between fact and fiction and ascertaining a clear intent to cause harm. Inadvertent errors or certain forms of opinion or belief, as well as expressions of satire and parody cannot easily be placed in a binary analysis of fact or fiction. Moreover, disinformation (with the intent to cause harm) can be spread online by innocent third parties with no intent to cause harm (which would be categorized as misinformation).

The main component that rings true is the *intent* of disinformation. The UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression in her recent report on "Disinformation and freedom of opinion and expression" emphasized that some forms of disinformation amount to incitement to hatred, discrimination and violence, which are prohibited under international law.[42]

Further references to regional definitions of disinformation are also elaborated in earlier policy brief papers by the Office of the OSCE Representative on Freedom of the Media.[43]

Ultimately, it is important to note that the right to freedom of expression is broad in its scope and not limited to "correct" statements. The right also protects "expression that may be regarded as deeply offensive,"[44] and that ideas, information and opinions "that offend, shock or disturb the State or any part of the population"[45] are also protected under the right to freedom

---

[39] Bukovska, OSCE RFOM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (2019); https://www.osce.org/files/f/documents/9/f/456319_0.pdf; Krivokapic, OSCE Non-Paper on the Impact of Artificial Intelligence on Freedom of Expression (2019); https://www.osce.org/representative-on-freedom-of-media/447829

[40] UN, OSCE, OAS, ACHPR, Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda (2017); FOM.GAL/3/17 (osce.org)

[41] A comprehensive overview of the challenges are provided in the OSCE RFOM Strategy Paper to Put a Spotlight on Artificial Intelligence and Freedom of Expression (2019); https://www.osce.org/files/f/documents/9/f/456319_0.pdf

[42] UNSR Irene Khan, Report on Disinformation and Freedom of Opinion and Expression (2021); A/HRC/47/25 - E - A/HRC/47/25 -Desktop (undocs.org)

[43] Expert roundtables on Disinformation | OSCE

[44] Human Rights Committee, General Comment No 34, CCPR/C/GC/34, 12 September 2011, para 11. 4

[45] Handyside v UK, Application No 5493/72, judgment of 7 December 1976 at para 49.

of expression. While, at the same time, this does not justify the dissemination of knowingly or recklessly false statements by official or State actors.[46]

Consequences for the OSCE's comprehensive concept of security

Human rights lie at the core of what the OSCE represents. From the very foundation of the CSCE/OSCE, its participating States aimed at conceptually innovating what "security" means.[47] The absence of conflict is one component; but equally relevant for security is the respect for human rights and fundamental freedoms, as is economic and environmental security. Ultimately, comprehensive security is about ensuring that people throughout the OSCE region are both *safe* and *free*.

The OSCE commitments provide a solid basis for comprehensive security; and while the OSCE would benefit from new commitments addressing emerging challenges in the digital context, existing commitments have stood the test of time and prove flexible enough to be applicable today.

When it comes to systemic societal issues like disinformation, content is not itself the problem. It becomes the problem, however, when it reaches a large audience, and particularly when it is combined with other, related content that reinforces the message.[48] AI tools linked to targeted advertising are one of the main enablers of widescale distribution of disinformation online. The focus should be on regulating dissemination and targeting techniques, as opposed to regulating content, which, as explained above, is often a flawed process, and may further deepen the problem. This is also very much in line with OSCE commitments pertaining to freedom of expression and media freedom.[49]

The automated selection of which information is made available and which content is suppressed in an online information space tailored to each individual, could have significant consequences on collective awareness of and engagement with politics, current affairs, and scientific consensus.

Democracy requires citizens to engage with opposing viewpoints, it also requires a shared value of and reliance on fact. Algorithmic gatekeeping tends to hold back opposing viewpoints that deter user engagement, while promoting sensational and deceptive content to increase engagement; meanwhile disinformation obliterates the notions of truth and fact. The perilous combination of targeted advertising and online disinformation not only weakens the exercise and enjoyment of individual human rights, but may erode the foundations of democracy, peace, security, and prosperity.

---

[46] UN, OSCE, OAS, ACHPR, Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda (2017); FOM.GAL/3/17 (osce.org)

[47] Zannier, Human Rights and OSCE's comprehensive security concept (2017); https://www.osce.org/files/f/documents/b/b/103964.pdf

[48] Cobbe and Singh, Regulating Recommending: Motivations, Consideration, and Principles (2019); https://ejlt.org/index.php/ejlt/article/view/686/979

[49] Already back in 1989, the OSCE participating States commited to "ensure that individuals can freely choose their sources of information" and that "to these ends they will remove any restrictions inconsistent with the above mentioned obligations and commitments." See https://www.osce.org/files/f/documents/4/f/99565_0.pdf