SPOTLIGHT ON **ARTIFICIAL INTELLIGENCE** AND **FREEDOM OF EXPRESSION**

A Policy Manual

osce
The Representative on Freedom of the Media

This publication is part of the project "Spotlight on Artificial Intelligence and Freedom of Expression" (#SAIFE).

6a Wallnerstrasse
1010 Vienna, Austria
Phone +43-1-514-36-68-00
e-mail: pm-fom@osce.org
https://www.osce.org/fom/ai-free-speech

ISBN: 978-92-9234-749-9

Spotlight on Artificial Intelligence and Freedom of Expression

A Policy Manual

Authors

Eliska Pirkova, Matthias Kettemann, Marlena Wisniak, Martin Scheinin, Emmi Bevensee, Katie Pentney, Lorna Woods, Lucien Heitz, Bojana Kostic, Krisztina Rozgonyi, Holli Sargeant, Julia Haas, and Vladan Joler

Editors

Deniz Wagner and Julia Haas

Experts

Jennifer Adams, Susie Alegre, Asha Allen, Andreas Marckmann Andreassen, Nikolett Aszodi, Jef Ausloos, Josephine Ballon, Joan Barata, Nadia Bellardi, Susan Benesch, Guy Berger, Frederik Zuiderveen Borghesius, Irina Borogan, Jonathan Bright, Elda Brogi, Amy Brouillette, Joanna J. Bryson, Pete Burnap, Camilla Bustani, Ignacio Talegon Campoamor, Maja Cappello, Marcelo Daher, Anita Danka, Nicholas Diakopoulos, Aijamal Djanybaeva, Leyla Dogruel, Maria Donde, Sead Dzigal, Franccesca Fanucci, Marc Fumagalli, Maximilian Gahntz, Jana Gajdošová, Maya Indira Ganesh, Lalya Gaye, Brandi Geurkink, Arzu Geybulla, Michele Gilman, Nadine Gogu, Gabrielle Guillemin, Rustam Gulov, Ben Hayes, Natali Helberger, Georgia Holmer, Andrea Huber, Karolina Iwańska, Sam Jeffers, Elliot Jones, Pascal Jürgens, Agnes Kaarlep, Frederike Kaltheuner, Kari Karppinen, Susan Kerr, Benjamin Kille, Yoojin Kim, Wolfgang Kleinwächter, Beata Klimkiewicz, Djordje Krivokapić, Ľuboš Kukliš, Andrey Kuleshov, Joanna Kulesza, Collin Kurre, Susan Landau, Paddy Leerssen, Emma Llansó, James MacLaren, João Carlos Magalhães, Samvel Martirosyan, Estelle Massé, Kyle Matthew, Eleonora Maria Mazzoli, Tarlach McGonagle, Marko Milosavljevič, Mira Milosevic, Iva Nenadić, Marielza Oliveira, Rebekah Overdorf, Roya Pakzad, Sejal Parmar, Patrick Penninckx, Jon Penny, Carlos Perez-Maestro, Emilija Petreska-Kamenjarova, Andrej Petrovski, Courtney Radsch, Otabek Rashidov, Judith Rauhofer, David Reichel, Moritz Riesewieck, Katitza Rodriguez, Asja Rokša-Zubčević, Bianca Schönberger, Christopher Schwartz, Lisa Seidl, Murtaza Shaikh, Jat Singh, Vanja Škorić, Andrei Soldatov, Maria Luisa Stasi, Nikolas Suzor, Damian Tambini, Dhanaraj Thakur, Gulnura Toralieva, Max van Drunen, Vitaly Vasilchenko, Francisco Vera, Kristina Voko, Diana Vlad Calcic, Ben Wagner, Douglas Wake, Hilary Watson, Agnieszka Wawrzyk, Veszna Wessenauer, and Andrej Zwitter

Observers

United Nations OHCHR, UNESCO, Council of Europe, European Audiovisual Observatory, European Commission, European Union Agency for Fundamental Rights, European Broadcasting Union, OSCE (Secretariat, ODIHR, HCNM)

Copy Editor

Tom Popper

Design and Layout

Peno Mishoyan

Table of Contents

Foreword

Dear Reader,

I am pleased to present my Office's publication putting a spotlight on artificial intelligence and freedom of expression (SAIFE). This publication is the culmination of two years of research and several expert workshops, bringing together the knowledge of more than one hundred of the most renowned scholars and practitioners working in the field of media freedom, human rights, technology, and security.

The year 2022 marks the 25th anniversary of the mandate of the OSCE Representative on Freedom of the Media. In 1997, the year this institution was established, only 1.7 percent of the global population was online, and digital technologies supporting online communication were novel and virtuously optimistic.

Twenty-five years on, the number of people who access the internet has risen to more than 80 percent across the OSCE region. This monumental expansion has been profoundly beneficial for freedom of expression, the free flow of information and the ability to seek, receive and impart information and ideas of all kinds across borders, and across the globe.

This has been crucial for economic, public and political participation, for democratization, for education and health, for holding power to account, and for shedding light on war crimes and other human rights violations. At the same time, it has also given rise to mass surveillance as well as cybercrimes, and the spread of illegal and harmful content online.

Managing the immeasurable plurality of information online has become impossible without the support of machine-learning technologies and other forms of artificial intelligence (AI). AI technologies are becoming the main tools for shaping and arbitrating content online; AI is used to decide on what content is taken down, what content is prioritized or to whom it is disseminated. These decisions are executed by technology that is developed and deployed by a handful of online platforms—the gatekeepers to the digital world.

These are powerful companies with the ability to shape and arbitrate political and public discourse. There is no doubt that the way online information is curated and moderated has a direct and significant impact on global peace, stability and comprehensive security. With such power must come responsibility. Nevertheless, these new gatekeepers – and their business practices—are developing at a rate that outpaces any legal or regulatory framework for the use of AI to shape our online information space.

We find ourselves at a crossroads.

The OSCE participating States must unite to find multilateral solutions for challenges to their common information space. They must do so by putting human rights at the centre of the development and deployment of AI for online content curation and content moderation.

These challenges are far reaching, and solutions can only be found through the action of many different stakeholders. With regard to the challenges pertaining to media freedom and freedom of expression, my hope is that this publication will assist OSCE participating States, policymakers, academia and media professionals throughout the region and beyond on how to collectively develop such human rights safeguards within their national, regional and international capacities.

December 2021

Teresa Ribeiro
OSCE Representative on Freedom of the Media

Key Recommendations for OSCE Participating States

1. Protect and promote freedom of expression and other human rights as the centre of AI-related strategies and policies
2. Preserve and foster the internet as a space for democratic participation and representation and for media pluralism
3. Develop evidence-based policies, built on inclusive processes, to respond to challenges to freedom of opinion, freedom of information and freedom of expression
4. Promote compliance with the UN Guiding Principles on Business and Human Rights, to prevent the prioritisation of profit maximisation at the expense of human rights and democratic values
5. Oblige online platforms to conduct human rights due diligence, including through human rights impact assessments (HRIAs) for their content governance policies and automated decision-making, as well as for their business practices, such as data harvesting, targeted advertising and interface design
6. Enforce clarity, explainability, and accessibility on the use of AI for content moderation, content curation and targeted advertising
7. Ensure that human rights protections are not fully outsourced or automated, and provide transparency about any public-private-partnerships

8. Enact strong transparency frameworks, including by mandating comprehensive transparency reports that contain detailed information on the use of AI

9. Make certain that robust remedy mechanisms against censorship and surveillance power are in place, including through human review and independent appeal mechanisms

10. Guarantee strong accountability, including through independent oversight and independent auditing, particularly of compliance with human rights and non-discrimination

11. Respect the right to privacy and data protection, including by identifying limits to surveillance-based advertising and by ensuring robust transparency and user agency in tracking- and profiling-based business practices

12. Promote media and digital literacy and strengthen users' empowerment, agency and control over content governance and the use of their data, including by providing the possibility to opt-out of all automated decision-making

13. Address unbalanced and monopolised market powers and promote plurality, and technological and media innovation

14. Engage on the multilateral level to ensure human rights s afeguards in the development and deployment of AI for online content curation and content moderation

Introduction: Upholding the Principles of the Helsinki Final Act in the Digital Age

Last year saw the 45th anniversary of the signature of the 1975 Helsinki Final Act. That Act, the outcome of the First CSCE Summit of Heads of State or Government, has become a cornerstone of Europe's political order. Eastern and Western states together agreed on ten principles that would guide their behaviour, including respect for sovereign equality and, in Principle VII, mutual respect for human rights and fundamental freedoms. The Helsinki Final Act further contains commitments on cooperation among states, including scientific and technological cooperation. Even computers feature in the Helsinki Final Act: cooperation was deemed necessary, especially regarding the development of "telecommunications and information systems; technology associated with computers and telecommunications, including their use for [...] automation, for the study of economic problems, in scientific research and for the collection, processing and dissemination of information".[1]

Cooperation and multilateral approaches are needed more than ever, and new actors shaping how information is processed, amplified, and curated necessitate new regulatory approaches to the human rights challenges of today's informational landscape. While states bear the primary obligation to respect, protect and fulfil human rights, internet intermediaries, and especially a few dominant social media platforms,[2] increasingly influence the realisation of these rights. On the internet, a new quasi-normative order that challenges traditional conceptions of normativity can be seen.[3] In today's digital world, the exercise of freedom of expression is increasingly governed in private, hybrid and public spaces that are shaped by private companies, states and users in different, highly asymmetric

1 Conference on Security and Co-Operation in Europe, Helsinki Final Act, https://www.osce.org/helsinki-final-act.
2 Online platforms fulfil a broad variety of functions, including to store and disseminate information. These include social media, search engines, ad networks and e-commerce platforms. This publication focuses on online platforms that are primarily characterised by facilitating interactions on the internet between persons by offering a communicative space. Some platforms primarily host and curate content, others additionally facilitate digital commerce. Platforms that primarily engage in facilitating communication, including for commercial purposes, are usually called social media platforms.
3 Matthias C. Kettemann, The Normative Order of the Internet (Oxford: OUP, 2020).

power relations. Moreover, these online ecosystems have paved the way for new forms of governance of expression, including those performed by algorithms and artificial intelligence (AI). More often than not, internet intermediaries' quasi-normative standards universally determine how free expression is governed, both in scope and intensity. This content governance is typically done outside of any public scrutiny and often performed by opaque automated decision-making at scale, with no guarantee of compliance with the international human rights framework.

The use of automation in content governance further exacerbates many existing challenges to human rights online, while giving rise to new ones. In general, AI tools are widely used for moderating and curating user-generated content as well as for delivering personalised ads. Automated and AI-based tools deployed in the moderation and curation of online content have been at the centre of academic and policy debate. Private actors and policy makers often present AI as a silver-bullet solution that already can, or a few years from now will be able to, resolve highly complex issues around the dissemination and distribution of potentially illegal or harmful content. However, the proactive and automated identification, detection and removal of online content carries systemic risks. Such AI-based tools are typically deployed by dominant private actors, and often required by states, either directly, through legally binding legislative frameworks, or indirectly, through increased pressure on intermediaries to "do more". In addition, the use of automation and AI to curate content, and thus promote some information at the expense of others, based on intermediaries' internal, profit-oriented policies, carries systemic risks as well. Several of these risks stem from automated decision-making systems directly linked to surveillance-based business models of very large internet intermediaries.

Many civil rights groups have been raising the alarm for years, pointing to ongoing human rights violations resulting from opaque automated decision-making. Internet intermediaries, such as social media platforms, have become essential for private communications and public discourse, and they are run by algorithms that determine people's access to information and thus process of opinion making. The predominant business models of the most powerful internet intermediaries are surveillance-based, often exploiting individuals' psychological vulnerabilities and other weaknesses. Built on a foundation of mass user-data collection and

analysis, these business models are part of a market ecosystem that Harvard Professor Shoshana Zuboff has labeled "surveillance capitalism".[4] Evidence suggests that surveillance-based business models have driven the distortion of our information environment in ways that are at odds with pluralism, diversity and democratic processes and decision-making. Recent revelations of whistleblower Frances Haugen only confirmed these allegations, underlining the need for states to establish a human rights-centred model of platform governance. Seeing a need to ensure protection of human rights, many call for increased state regulation. However, regulating internet intermediaries, especially regulating their use of AI and algorithmic systems with the goal of mitigating their societal risks, is challenging and multifaceted.

In general, very few examples of good practice of human rights-compliant content governance exist, and some voluntary commitments to improve the protection of human rights pledged by internet intermediaries have ultimately proven insufficient. The time has come, therefore, to move towards principles for human rights-centred online ecosystems, and, in this regard, to contribute to upholding the principles of the Helsinki Final Act in the digital age. Such an Act could again unite those seeing the internet as an extension of their national borders and those wishing to pursue more human rights-focused policies. It would thus reaffirm the very foundation of the OSCE: that human rights are an integral part of its comprehensive security, online and offline.

The goal of the Spotlight on Artificial Intelligence & Freedom of Expression project (SAIFE) is to provide guidance to OSCE participating States on how to fulfil their positive obligation to protect human rights of individuals when creating regulatory responses to the new challenges facing the right to freedom of expression in the digital age. Four expert workshops were organised to identify the actual and foreseeable negative impact that automated and AI-based methods for detecting, evaluating, curating and personalising online content have on individuals' human

4 Ranking Digital Rights, It's the Business Model: How Big Tech's Profit Machine is Distorting the Public Sphere and Threatening Democracy (2021).

rights. The expert workshops put an emphasis on the individual right to freedom of expression and opinion, as well as rights on a societal level, including media freedom. The workshops resulted in a set of human rights-centred recommendations with the aim of identifying human rights due diligence measures, and procedural safeguards to address both individual and societal risks arising from an unwarranted use of AI in content governance.

Structure and Executive Summary

In the framework of the SAIFE project, the Office of the OSCE Representative on Freedom of Expression, together with Access Now, organised four expert workshops in the first half of 2021. These expert workshops unpacked and analysed the main challenges that AI tools pose to human rights, in particular, the right to freedom of expression and opinion, and media freedom and pluralism. The workshops focused on four main thematic issues:

- Content Moderation—Security
  AI-based tools deployed in content moderation to detect and evaluate illegal content online, including security threats, such as extremist and terrorist content.

- Content Moderation—Hate Speech
  AI-based tools used for detecting and evaluating potentially harmful but legal content, with a specific focus on online hate speech and algorithmic discriminatory bias.

- Content Moderation—Media Pluralism
  AI-based tools designed for curating and personalising online content, with a focus on content recommender systems and their impact on media pluralism.

- Content Moderation—Surveillance
  AI-based tools used in surveillance-based advertisement and their link to curating content through profiling of individuals and predicting future behaviours.

This report contains the main findings of these expert workshops as well as policy recommendations addressed to the OSCE participating States, while acknowledging that a multi-stakeholder approach is needed to effectively and sustainably address the complex challenges that content moderation and content curation pose to freedom of expression. The recommendations for OSCE participating States were put forward during the workshops and reviewed by renowned experts in the field of freedom of expression, media pluralism and artificial intelligence. The publication is based on outcome reports of each expert workshop. The outcome reports were co-drafted by the Chair appointed to lead the work of the individual expert groups, the rapporteurs of the respective expert workshop and Eliška Pírková of the project's Implementing Partner, Access Now, and involved consultation with all expert and observer participants of the respective workshops. The report follows the structure of the thematic areas addressed by each expert group. It can be viewed as four separate blocks, each providing human rights-centred policy recommendations addressed to the OSCE participating States.

AI in content moderation

The outcomes of the first two expert workshops, focusing on the use of AI in content moderation to target illegal content and potentially harmful content, such as hate speech, have been merged into one joint section. The section provides a set of policy recommendations intended to help prevent the negative impact that AI tools in content moderation have on the right to seek, receive and impart information and ideas of all kinds.

Content Moderation—Security

AI-based tools deployed in content moderation to detect and evaluate illegal content online, including security threats such as extremist and terrorist content

One of the two working groups looking at content moderation focused on automated and AI-based systems used to detect and act upon illegal content and accounts associated with spreading such content. This practice includes filtering and hash-matching technologies deployed to block uploads, and tools to take down or de-rank content *ex post*, often with cross-border effect. Notable challenges emerge when AI

technologies are used to monitor national law or even to allow for the monitoring by law enforcement of peoples' digital communications under the justification of security and public safety. Individual and group anonymity can be under special pressure, which may lead to chilling effects on freedom of expression and freedom of the media, as well as the safety of journalists. While the impact of AI-based content moderation on illegal conduct remains unclear, AI technologies are context-blind and are prone to overbroad application of the rules they seek to impose. This means that they regularly generate so-called false positives and false negatives in identifying presumably illegal content online. The result can be arbitrary restrictions of legitimate expressions or failure to restrict illegal expression.

The working group outlined the potential negative impact that using AI-based tools in content moderation has on individuals' freedom of expression, and the wider societal risks they pose for freedom of the media, democracy and the rule of law. The policy recommendations set forward by the working group investigating content moderation and illegal content enable OSCE participating States to identify, analyse and assess significant systemic risks stemming from content moderation systems, including when they are used to prevent the rapid dissemination of illegal content online. These recommendations are combined with those of the working group on legal but harmful content, including hate speech, to provide recommendations on free speech safeguards for AI in content moderation, as well as guidance for transparency, data access, independent oversight, remedies and frameworks for human rights due diligence.

The work of this expert group and development of this part of the report was led by the Chair, Prof. Martin Scheinin, and supported by rapporteurs Prof. Matthias Kettemann and Marlena Wisniak.

Content Moderation—Hate Speech

AI-based tools used for detecting and evaluating potentially harmful but legal content, with a specific focus on online hate speech and algorithmic discriminatory bias

The second working group on content moderation addressed the actual and foreseeable negative impact that automated and AI-based tools for detecting and evaluating online hate speech have on individuals' human rights, with an emphasis on marginalised groups' right to freedom of expression and opinion. The impact of discriminatory bias can manifest as "biased censorship" against content posted by members of specific societal groups that are often targeted by hateful expressions and abuse online. While hate speech itself is highly context-dependent and difficult to detect and remove automatically, groups likely to be targeted by online abuse may be silenced as their own communications are censored. Datasets are used to train automated tools to identify and distinguish different categories of content. If these datasets do not include examples of speech in different languages and from different communities, or if certain groups are not represented in the training data, this can lead to erroneous classifications that disproportionately affect marginalised groups. Automated tools may either miss potentially hateful content (false negatives) or wrongfully label legitimate expressions as hate speech (false positives).

The joint recommendation on free speech safeguards for AI in content moderation provide guidance for transparency, data access, independent oversight, remedies and frameworks for human rights due diligence. The specific recommendations from the "hate speech" working group aim at enabling the identification and addressing of systemic risks, especially for marginalised groups, stemming from AI-based content moderation systems deployed to detect potentially harmful content, such as hate speech. These recommendations provide guidance on human rights-friendly automated content moderation tools and on increasing the digital participation of marginalised groups in public discourse.

The work of this expert group and development of this part of the report was led by the Chair, Prof. Lorna Woods, and supported by rapporteurs, Emmi Bevensee and Katie Pentney.

AI in content curation

Content Curation—Media Pluralism

AI-based tools designed for curating and personalising online content, with a focus on content recommender systems and their impact on media pluralism

The first part of the section on content curation analyses the negative impact of algorithmic content recommender systems on individuals' human rights, with an emphasis on the absolute right to freedom of opinion as well as media pluralism and media freedom. It addresses: the amplification of potentially harmful content, such as deceptive, polarising or hateful content; the impact of recommender systems on diversity of opinions and ideas; the impact of algorithmic curation on the right to form an opinion and media plurality; and the risk of polarisation of societies. The algorithmic selection of content is based on intermediaries' policies, which follow their internal and advertisers' economic interests rather than focusing on accuracy, diversity or public interest (such as news value). This approach affects public communication and the free flow of information, while putting pressure on professional journalism by channelling advertising money to intermediaries. Moreover, news items are accessed less often than a bundled overall offer of individual information content, so that every single post fights for attention in the news feed, which encourages the use of clickbait to engage users. While this model facilitates advertising and generates profit for intermediaries, it poses a challenge for media pluralism.

After describing the challenges, this part puts forward a set of policy recommendations for OSCE participating States to ensure meaningful transparency by internet intermediaries, increased individual agency and control, along with recommendations to promote diversity of voices, public interest information and media pluralism.

The work of this expert group and development of the report was led by the Chair, Prof. Krisztina Rozgonyi, and supported by rapporteurs Lucien Heitz and Bojana Kostic.

Content Curation—Surveillance

AI tools used in surveillance-based advertisement and their link to curated content through profiling of individuals and predicting future behaviours

The second part of the section on content curation focuses on the nexus between content curation and advertising. AI in targeted advertising refers to the practice of directing specific advertisements at individuals based on the use of automated statistics—e.g. machine learning, natural language processing, speech recognition, and image recognition. Various forms of data exploitation, including psychological profiling and nano-targeting, are enabled by the processing of data, signal extraction and automated analysis of a wide variety of different types of data—such as user-generated content, location data, behavioural patterns, psychographics, information about the user's race, economic status, sex, age, generation, level of education, income level and employment. The short and long-term as well as direct and indirect effects of this surveillance-based advertising on human behaviour, well-being, and society in general are not yet known, but AI-based systems have repeatedly produced biased and erroneous outputs.

This part of the report analyses the far-reaching impacts that automated and AI-based processes used for surveillance-based advertisement have on individuals' personal interactions, communication, and participation in democratic debates. From privacy violations to fragmentation of informational spaces, surveillance-based advertisement may seriously harm the right to freely form and hold opinions, as well as to seek, receive and impart information. The working group producing this part of the report tackles issues such as: the inherent lack of explainability and transparency of algorithmic systems fed with individuals' personal and behavioural data; manipulative marketing practices exploiting particular characteristics and users' vulnerabilities in order to increase the persuasiveness of a message; discrimination caused by algorithms optimising advertising; and amplification of potentially harmful content in order to increase users' engagement, with a view to enhancing profit.

Recommendations based on this analysis include measures intended to increase transparency and to prevent and mitigate the human rights risks stemming from practices such as intrusive targeting and personalisation of content. The recommendations also underline the need to tackle surveillance-based business models of a few dominant internet intermediaries. Policy recommendations to OSCE participating

States include empowering individuals to exercise control over their data and the information they receive and impart, as well as better protection of freedom of opinion in the digital ecosystem.

The work of the expert group and development of the report was led by the Chair, Prof. Vladan Joler, and supported by rapporteurs Holli Sargeant and Julia Haas.

AI in Content Moderation

AI in Content Moderation with a Particular Focus on Security Threats and Hate Speech

This part of the report focuses on the use of AI in content moderation, and the human rights implications stemming from the use of AI tools to target specific categories of user-generated content. It highlights shortcomings of AI in content moderation in the context of both manifestly illegal content, such as terroristic or extremist content, as well as potentially harmful, yet legal content, such as hate speech, in particular from the perspective of marginalised communities. It concludes by providing operational and technical human rights-centred recommendations for OSCE participating States. These recommendations are intended to address the existing negative impact of AI tools in content moderation on the right to seek, receive and impart information and ideas of all kinds.

1. Defining the scope of content moderation

Two expert workshops focused on the use of AI tools in content moderation, primarily addressing two categories of user-generated online content: illegal content and potentially harmful but legal content, with specific emphasis on hate speech. The following sections explain the scope of the expert groups' work in each area.

1.1 Security threats and illegal content online

Automated detection tools aimed at potentially illegal content online—also referred to as proactive measures—have been in the centre of academic and policy debate. Private actors and policymakers often present AI as a silver-bullet solution that will eventually be able to resolve highly complex issues around the dissemination of illegal content, including the spread of terrorist propaganda. However, this view of the technology, which is presented as justification to boost "AI uptake across the economy, both by the private and public sector"[5] disregards the systemic risks involved

5 European Commission, Annex to the Communication from the Commission to the European Parliament, the European Council, the Council, the European economic and Social Committee and the Committee of the Regions, "Coordinated Plan on Artificial Intelligence" (7 December 2018, COM(2018) 795 final, p. 4, at <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56017>.

in proactive identification, detection and removal of user-generated content. While addressing security threats is legitimate and necessary, responses must not come at the expense of human rights. Risks stem from the automated decision-making systems that are deployed by online platforms and are often imposed by states, either directly, through legally binding legislative frameworks, or indirectly, through increased pressure on platforms to "do more".

Regardless of the specific technological method used, such automated tools may impose prior restraints on the right to freedom of expression and information. In practice, this means that they may a priori exclude certain persons, groups, ideas, or means of expression from public discourse. There are strict requirements for justifying prior restrictions of freedom of expression in the international human rights framework and in various constitutional laws. These requirements stem from concerns about overly restricting the free flow of information. In this regard, AI tools, are especially worrying, because these systems are shielded from any public scrutiny, are context-blind and operate in a highly non-transparent manner that prevents any possibility of effective remedy and redress. While prescreening content to limit the spread of malware and child sexual abuse has been broadly accepted as a positive use of automation, one has to remain cautious about applying the same logic to other types of speech that fall into a broader area of content governance.[6]

Given that a large number of legislative proposals to regulate potentially illegal online content have recently been introduced across the OSCE region, the work of the expert group is particularly significant. The group's outcome report provides human rights-centred recommendations for better regulation of AI tools in content moderation. It is intended to help identify rights-respecting regulatory responses to the spread and dissemination of illegal content online.

While this outcome report does not define what constitutes potentially illegal content, its security-focused recommendations look at proactive methods for detecting and evaluating:

6 Emma Llanso, "No amount of 'AI' in content moderation will solve filtering's prior-restraint problem" Big Data & Society 7(1), pp 1-2, at <https://journals.sagepub.com/doi/pdf/10.1177/2053951720920686>.

- **Content that is illegal irrespective of its context**
  A typical example of such content is child sexual abuse, which is prohibited by a number of international legal instruments, such as the Council of Europe Budapest Convention, the Lanzarote Convention, Convention 182 of the International Labour Organization, the UN Convention on the Rights of the Child and others. However, even for this content category, national laws do not provide a uniform response.

- **Content that is a part of a wider crime**
  For instance, in the case of beheading videos that go viral, at least one violent crime has taken place in "real life". Any content moderation initiative that fails to take the offline elements of a crime into account risks leaving victims without redress. Furthermore, such online content, as well as its removal, can have an impact on investigations (as evidence) and documentation of human rights abuses.

- **Legal content that is illegal due to its context**
  This refers to content that is not in itself illegal, but the manner in which it becomes available online can amount to a criminal offence. A typical example of such content is the depiction of non-consensual nudity or unauthorised publication of personal information.

- **Content that is illegal mainly due to its intent and effect**
  This category includes incitement to violence or incitement to terrorism. Usually, it is not the content itself, but rather the (subjective) intent behind its publication, coupled with the (objective) risk that some recipients will be incited to violence, that constitutes an offence. This category also includes, for example, xenophobia, incitement to discrimination and incitement to hatred.

1.2 Hate speech online

The large number of internet intermediaries of various shapes and sizes has created a global marketplace of ideas, enabling individuals across the world to share and receive information and ideas. At the same time, however, it has also enabled the proliferation and amplification of hate speech.[7] States must grapple with the competing interests of protecting free speech of individuals while simultaneously upholding the rights and freedoms of the targets and recipients of hate speech, as well as the public at large. In particular, the exercise of human rights can be curtailed for marginalised groups, who are subject to discriminatory bias and are often silenced by societal phenomena such as hate speech. The manifestation of hatred is not unique to the online context. On the contrary, it has existed in the "real world" and targeted individuals and groups on the basis of identifiable characteristics, such as race, sex/gender, religion and sexual orientation across societies and throughout history. However, the online dimension presents new challenges in terms of the volume, reach and impacts of hate speech. For instance, Facebook removed more than 20 million pieces of hate speech content in the last quarter of 2020 alone,[8] while Google removed nearly 100,000 videos from YouTube during the same period.[9] Networked hate speech can also prove more difficult for its targets (or the public at large) to avoid or tune out, as speakers can reach into traditionally safe spaces, including people's homes, often under the veil of anonymity, and in some cases through coordinated smear campaigns.[10]

In response to this growing phenomenon, and in the wake of concerns about its societal impacts, efforts to combat hate speech more effectively have significantly increased over the last few years. There has been a focus

7 See, e.g., European Commission, Countering illegal hate speech online: 5th evaluation of the Code of Conduct (June 2020) at <https://ec.europa.eu/info/sites/default/files/codeof-conduct_2020_factsheet_12.pdf>.
8 Facebook Transparency, Community Standards Enforcement Report at <https://transparency.facebook.com/community-standards-enforcement#hate-speech>.
9 Google Transparency Report, "Featured policies: Hate Speech" (Oct 2020 —Dec 2020) at<https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en>.
10 Matthew Williams and Mischon de Reya, "Hatred Behind the Screens: A Report on the Rise of Online Hate Speech" (2019) p. 18 at <https://hatelab.net/wp-content/uploads/2019/11/Hatred-Behind-the-Screens.pdf>.

on questions of how best to combat hate speech in the online environment, the differentiated roles and responsibilities of states and private actors in moderating hate speech online, and the role to be played by automated decision-making systems in detecting and removing such content.

At the same time, there is no universally accepted definition of hate speech at the international level. This lack of definition has left room for courts and tribunals to determine the boundaries of permissible and impermissible expression. A wide range of expression may fall within the scope of hate speech: from illegal hate speech, such as incitements to violence of genocide, at the most severe end of the spectrum; through to potentially unlawful hate speech, such as threats of violence and harassment; to speech that does not reach the threshold of illegality but is nonetheless harmful and offensive.[11] With the proliferation of online platforms and user-generated content, the task of defining and regulating hate speech has increasingly been delegated to private companies. However, states retain ultimate responsibility for safeguarding human rights—including freedom of expression, non-discrimination and access to appropriate remedies. It is of the utmost importance that appropriate guidance is provided and oversight is ensured when private corporations intervene in the digital marketplace of ideas.

Freedom of expression requires protecting not only information and ideas that are received favourably, but also those that offend, shock or disturb.[12] Any restriction to the right must be legitimate, proportionate, and in accordance with international law. While cases of removal of illegal content may be clear; the position is more complex for content that does not meet the threshold of illegality but is harmful, and might interfere with the rights of others. It is challenging to define this second category of content, and to identify appropriate responses to it. Historically marginalised groups within society, whose voices are often not heard and who may not be represented in the halls of power, are frequent targets of hate speech. It is therefore critical to ensure greater participation of marginalised individuals and groups in decision-making around these

11 UN Strategy and Plan of Action on Hate Speech (2020), Table 1, p. 16 at <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf>.
12 Handyside v United Kingdom, App no 5493/72 (ECHR, 7 December 1976) [49]; UN Strategy and Plan of Action, p. 14.

fundamental questions, including discussions on how to effectively address hate speech. In practical terms, the absence of participatory and representative decision-making has led to overbroad and under-inclusive approaches to hate speech, particularly in the online environment.[13]

A lack of understanding, and lack of inclusion, can lead to situations where free expression of marginalised communities are improperly labelled as hate speech, allowing automated decisions to effectively silence individuals and groups. This may arise from a misunderstanding of context, including in-group and out-group dynamics. For instance, the terms "queer" and "gay" may be used as homophobic or transphobic slurs, defined and regulated as hate speech; but they may equally be reclamations by members of the LGBTQ+ community, or used for "pro-social functions", such as building communities and in-groups, and helping individuals to better prepare for and cope with hostility.[14] Similar misunderstandings of context and intent have been shown to result in over-removals of racial minorities' content online.[15] The regulation of hate speech is necessarily a contextual exercise—from the intention of the speaker, to the likely effects of the speech, to the particular meaning of the words or images in the given sociopolitical context. Studies have shown that automated decision-making is simply not capable of this contextual exercise. Blanket or over-inclusive approaches, therefore, may result in censoring members of marginalised groups, in violation of their freedom of expression.[16] This silencing effect should be a foremost concern for states and internet intermediaries alike.

---

13 See e.g. Molly K. Land and Rebecca J. Hamilton, "Beyond Takedown: Expanding the Toolkit for Responding to Online Hate" in Predrag Dojcinovic (ed.) Propaganda, War Crimes Trials and International Law: From Cognition to Criminality 143 (Routledge, 2020), p. 2 at <https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3514234_code858831.pdf?abstractid=3514234&mirid=1>.

14 Thiago Dias Oliva, "Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online" (2021) Sexuality & Culture 25, pp. 705-7 at <https://www.researchgate.net/publication/345501707_Fighting_Hate_Speech_Silencing_Drag_Queens_Artificial_Intelligence_in_Content_Moderation_and_Risks_to_LGBTQ_Voices_Online>

15 Thomas Davidson, Debasmita Bhattacharya and Ingmar Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets" (2019) at <https://www.aclweb.org/anthology/W19-3504.pdf>; Maarten Sap et al, "The Risk of Racial Bias in Hate Speech Detection" (2019) at <https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf>.

16 ibid.

Where hate speech policies are under-inclusive—that is, where they fail to address speech that is lawful but harmful—online spaces may become an unsafe or unwelcome environment for members of marginalised groups, effectively pushing them out. This is particularly problematic in light of the important role played by these online environments in our new (digital) marketplace of ideas, where individuals increasingly turn to share ideas, consume news, and participate in public debate. The result may be a "democratic deficit", whereby individuals from marginalised groups—women and non-binary persons, racial and ethnic minorities, members of the LGBTQ+ community, etc.—are unable or unwilling to fully participate in the democratic discourse.[17] Moreover, policies may be underinclusive in failing to account for intersectionality—that is, hate speech targeting individuals or groups on the basis of two or more identifying factors.[18]

With respect to the hate speech-focused recommendations, the following comments concerning scope should be borne in mind:

- While hate speech is not defined, this report focuses on lawful, but harmful, hate speech in the online environment. The recommendations are tailored to the regulation and moderation of such speech in a human rights-compliant manner.

- In light of the disproportionate impacts of hate speech moderation on marginalised communities, the report provides tailored recommendations for OSCE participating States to ensure protections for marginalised individuals and groups—including their right to freedom of expression, non-discrimination and access to adequate remedies.

- While content moderation is occurring at varying levels—as addressed in more detail below—this report primarily focuses on large-scale or "industrial" moderation of hate speech, to reflect the scale and extent of its impacts on freedom of expression.

17 Nani Jansen Reventlow, "The power of social media platforms: who gets to have their say online?" Lilith (4 February 2021) at <https://www.lilithmag.nl/blog/2021/2/3/the-power-of-social-media-platforms-who-gets-to-have-their-say-online>.
18 UN Strategy and Plan of Action, p 28.

2. Guiding note on content moderation

The detection and moderation of content that is either illegal, or potentially harmful but lawful, is a difficult task from beginning to end. While it is critical to hold internet intermediaries to account, understanding the limits of the technology itself (as well as the business models involved) helps public authorities to take more impactful action regarding corporations in the social media sector. This section provides an overview of algorithmic content moderation tools and techniques before identifying several points of vulnerability in content moderation at scale.

When using AI tools for content moderation, adequate justification and rationale of decisions is frequently lacking. This means that users often do not know why an automated decision was taken and what specific information was input that lead a machine to make a specific decision.

Types of content moderation

There are three basic models of content moderation, as defined by Robyn Caplan:

- Artisinal
  Content moderation by small in-house teams of human moderators.

- Community reliant
  Content moderation predominantly relying on volunteer community moderators from various subsections of the intermediary, as used by Wikipedia or Reddit.

- Industrial
  Content moderation involving large-scale outsourcing of moderation by humans combined with proprietary automated machine-learning detection.[19]

19 Robyn Caplan, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches (2018), at <https://datasociety.net/library/content-or-context-moderation/>

This report primarily focuses on the impacts and considerations of the third type of moderation, "industrial".

Content moderation tools can be broken down into the following categories:

- **Detection**
  Locating and identifying content that may violate an internet intermediary's policies.

- **Adjudication**
  Determining if detected content actually is in violation of an intermediary's policy.

- **Enforcement**
  Acting on content based on the consequences outlined in the intermediary's policy.

- **Appeal**
  Returning to the adjudication stage if a user contests or appeals an intermediary's judgment.

- **Policy**
  The set of principles, rules or guidelines that determine what content is acceptable on an intermediary's platform. In practice, these guidelines are reviewed and updated based on other components of the content moderation process.[20]

When content is identified as either being illegal or violating an intermediary's terms of service—or when it is predicted to fall into one of these categories—there are several possible outcomes. The most common are flagging or deletion. In case of deletion, the content is immediately removed and sometimes prevented from being uploaded again. Beyond taking down content, there are a number of ex post tools available for addressing "problematic" content. Some examples include:

20 Meedan, Content Moderation Toolkit: Toolkit for Civil Society and Moderation Inventory, at <https://meedan.com/reports/toolkit-for-civil-society-and-moderation-inventory/>.

- **Content demonetisation**
  On platforms like YouTube and Twitch, where creators can profit from their contents' popularity, Terms of Service can be enforced in a way to disable users from profiting from specific types of content. While such demonetisation may have advantages, it is often disproportionately enforced against marginalised people, either because of the above-mentioned challenges of the algorithm or for reasons of intentional silencing.

- **Content deprioritisation and deranking**
  What a user sees on online platforms is generally controlled by a range of private algorithms designed to increase engagement. Internet intermediaries can derank or remove the prominence of offensive or harmful accounts. While this can be useful for confronting the reach of harmful content, such as hate-speech or misinformation, it often devolves into an upranking of already popular content, such as mainstream news, potentially at the expense of marginalised voices.

- **Account suspension or feature limiting**
  Temporary suspensions provide a disincentive for users to violate community guidelines without permanently banning them. While such suspensions can prevent additional hate speech, they may also be used against marginalised people who are attempting to create a safe space in their corner of the internet.

- **Account removal**
  Removal of an entire account can disrupt a user's ability to maintain a large following base and can hence be an especially impactful response for serial offenders. However, as recent takedowns on Telegram have shown, extremist communities adapt quickly to recreate channels and followings in the wake of removals.

- **Block/Mute/Unfriend**
  These options provide a form of subjective moderation that allows users to choose which content they do not want to see on their personal feed. On social media platforms associated with Secure-Scuttlebutt, blockings are transparent in a way that telegraphs trust and distrust as a tool for limiting the spread of undesired messages.[21]

21 For further details, please see Scuttlebutt social network, a decentralised platform, at <https://scuttlebutt.nz/>.

Industrial algorithmic content moderation

Algorithmic content moderation involves a range of techniques from statistics and computer science that vary in complexity and effectiveness. All these techniques are designed to identify, match, predict, or classify user-generated content on the basis of its exact properties or general features. Automated tools are deployed by internet intermediaries to police content at scale across an array of issues, including terrorism, graphic violence, "toxic speech", non-consensual nudity, child abuse and spam detection. Two types of algorithmic content moderation are used predominantly, though not exclusively, to combat potentially illegal content online: text analysis and image analysis. When a certain piece of content is flagged by AI tools as potentially illegal, it is then typically placed in a queue, or prioritised, to be reviewed by a human "expert" moderator. It can then be deleted, or addressed using one of the ex post tools mentioned above.

Machine-learning systems that conduct text analyses regularly deploy natural language processing (NLP). NLP systems parse text in a comprehensive manner, attempting to bring the analysis closer to human understanding of the text in question. NLP tools are trained to predict whether specific text conveys positive or negative emotions (so-called sentiment analysis), and consequently, to classify whether it belongs or does not belong to a certain category of user-generated content. NLP is designed to predict outcomes based on labelled instances, for instance "offensive" or "not offensive". The best-known example of an NLP tool is Google/Jigsaw's Perspective API, an open-source toolkit that allows website operators, researchers, and others to use their machine-learning models to evaluate the "toxicity" of a post or comment.

Automated detection and identification of images and videos, on the other hand, often involves detecting content that was previously identified as illegal, while also discovering novel content that could be added to the category of illegal. Image detection and identification technologies use so-called hash values. A hash is a unique numerical value, also referred to as a "digital fingerprint", that is generated by a specific algorithm run on an image file. Simple hashing technology evaluates the dimension of the image or color values of pixels. A simple alteration of an image's pixel completely changes the hash of such a file, which means the tool is easy to circumvent. More nuanced tools use perceptual hashing that

includes fingerprints of images and videos mixed with other features of the content, such as hertz-frequency over time in audio, for example. Perceptual hashes are more robust and can identify images and videos even after their alteration. A typical example of perceptual hashing is PhotoDNA, which is developed by Microsoft and is used to combat child abuse online.

In the aftermath of the Christchurch atrocity in New Zealand in 2019, Facebook, Google, Twitter and Microsoft created the Global Internet Forum to Counter Terrorism (GIFCT), an organisation founded as part of their commitment to increase companies' voluntary compliance with the EU Code of Conduct to combat illegal hate speech online. Within the GIFCT framework, the four companies share best practices for developing their algorithmic content moderation tools. They also operate a highly secretive and non-transparent hash database of terrorist content, in which they share with one another digital fingerprints of "illicit content", including images, video, audio and text. For example, within hours of the Christchurch attack, Facebook uploaded hashes of approximately 800 different versions of the shooter's video. In theory, every single video uploaded by Facebook, YouTube and Twitter users can now be hashed and checked against the database. Content that matches an entry in the database will be immediately blocked. The database is solely operated by private actors and outside the realm of any public scrutiny, leading to severe challenges for journalistic and artistic content.

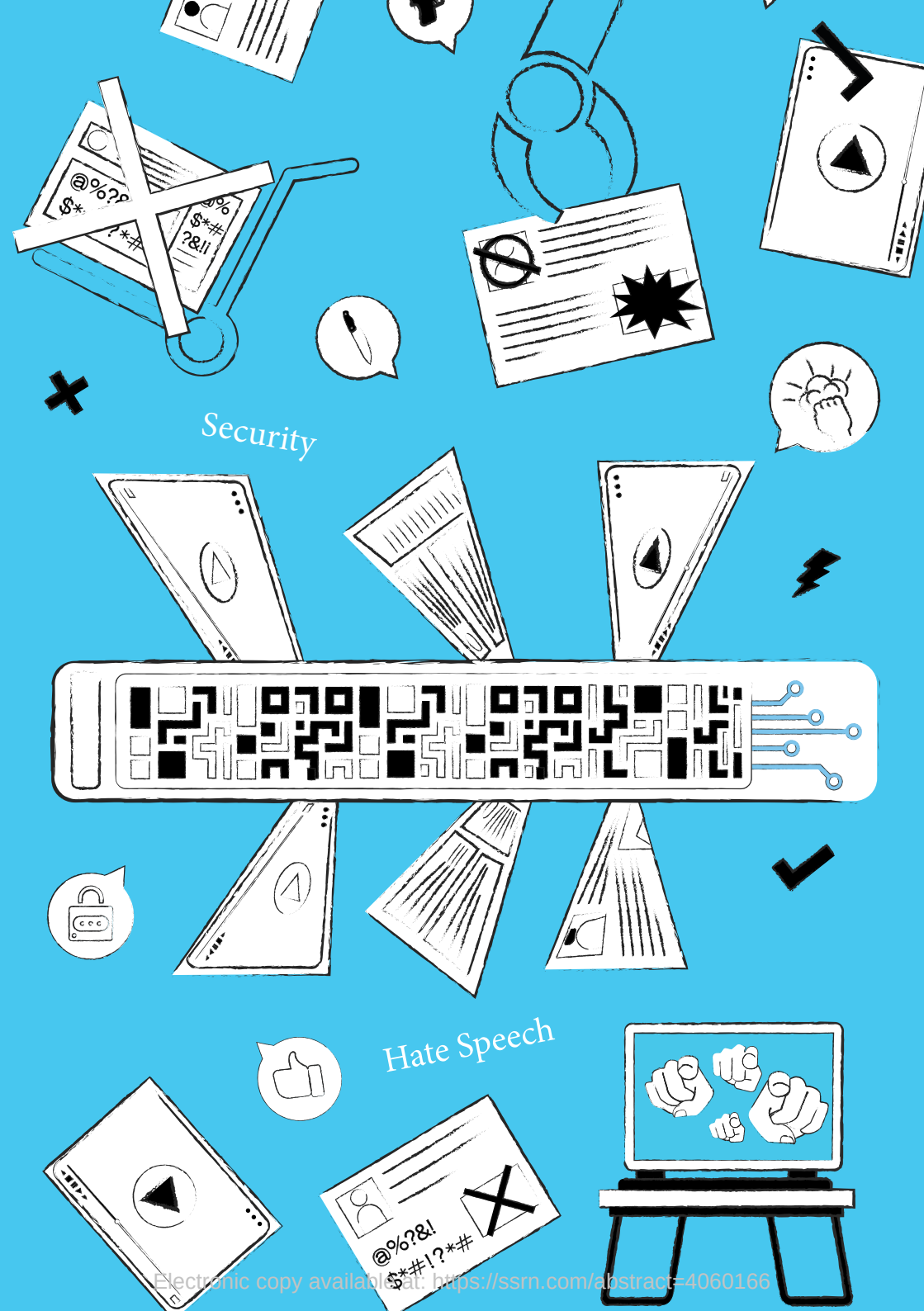
Shortcomings of algorithmic content moderation

There are several points of vulnerability in the design and development of algorithmic content moderation tools—as well as in the business models of internet intermediaries that use those tools —and these should be borne in mind by policymakers. One major vulnerability in the design of a machine-learning algorithm arises when a team of humans decides the rules for annotating the training data that will be used in the machine-learning model. This step is critical because AI is basically just a copy-machine. AI systems learn what humans teach them to learn—and even then, there might still be some deviation. The biases of the humans involved and those embedded in the data itself will replicate throughout the lifecycle of the AI system.

For instance, where a system attempting to detect crime online relies on structurally racist data, it will more deeply entrench racist outputs. Additional challenges arise from subtleties of speech, or the above-mentioned in-group and out-group usage of a given term, which could lead to systematic mislabeling of terms, causing harm to those a system is meant to protect. There is no single easy AI fix to online hatred, because each identity-based type of hatred and context is different and because the landscape is constantly changing as adversaries adapt. The environment in which machine-learning systems are created and deployed, especially for something as delicate as hate-speech detection, is deeply dynamic and contextual.

While transparency and participatory processes to address these challenges could significantly mitigate their risks, internet intermediaries have private interests that conflict with the kind of transparency required. For example, a machine-learning algorithm for detecting hate speech is itself a commodity that can be sold. As such, intermediaries are likely to make it proprietary. Moreover, many intermediaries claim, with whatever degree of truth, that sharing such algorithms would enable adversaries to abuse them. Furthermore, it can be difficult to explain the specific types of decisions for which an AI tool is employed, or to ensure that an AI tool generalises to address new forms of a problem it had been deployed to solve.

A robust content moderation framework should ensure that responses to illegal or potentially harmful content are proportionate and accurate, while seeking to address the technical and socio-political issues surrounding moderation. While states generally have indirect control over the moderation practices of corporations, full awareness of the policies and practices in place, as well as alternative possibilities, helps to inform and guide policy making.



3. Human rights-centred recommendations on the use of AI in content moderation

3.1 Recommendations on transparency

Recommendations for algorithmic transparency

- **States should oblige internet intermediaries to provide documentation on the AI tools they deploy for content moderation.** Any disclosure should be understandable and accessible for all users. Platforms should disclose what inferences are drawn about users' personal protected characteristics (i.e. age, race, gender, disability) or their associations, community memberships and proxies. Platforms should share information related to:

    - Training data: the content and origin of datasets used for training algorithms; methods for training AI models; variables/features/characteristics that influence algorithmic content curation; recommendation and/or ranking systems (e.g. users' age, gender, etc.) and how much control users have over those variables; and processes around management of training data (e.g. collection, storage, pre-processing/processing, transferring, retention).

    - Data enrichment services: data preparation and cleaning (data annotation/labelling, sentiment analysis, image recognition, speech to text validation, etc.) and "human-in-the-loop" tasks (human content moderation, developing a continuous feedback loop, validating algorithmic outputs and models, etc.)—including documenting the person(s) conducting the data enrichment service and information about their training.

    - The processes and results of testing, evaluating and validating these models, including quality and accuracy measurements.

- **States should mandate documentation of content-specific models by internet intermediaries.** The intermediaries should be legally mandated to disclose the criteria, parameters and features used for machine-learning models intended for content curation, content

moderation, and any other data analysis or pattern recognition. This should include disaggregating data for machine-learning models designed for taking down and removing user-generated content and for models designed to amplify and de-amplify "shadow banning" and deranking content. Any disclosure should be understandable and accessible for all users, while ensuring their privacy and data protection.

- **States should ensure diverse datasets, based on diverse attributes, as only attributes that are measured and recorded can be included in training or evaluation data for an algorithm.** Many widely available datasets focus on immutable characteristics (such as ethnic groups) or characteristics that are recorded and regulated by governments (such as legal gender, monetary income or profession). In contrast, characteristics like sexual orientation and gender identity are frequently not observed. This is a serious challenge for combating intersectional discriminatory bias inherent to some algorithmic systems.

- **States should implement transparent and human rights-centred use of AI systems by the public sector, including the use of AI tools for content moderation.** States should establish mechanisms for elevated scrutiny and transparency requirements when the public sector uses AI systems for content analysis—such as for facial recognition technologies and monitoring of content shared on online platforms.

- **States should oblige internet intermediaries to notify users when they are subjected to automated processes and when automated systems are used to moderate third-party content, and platforms should explain how such mechanisms operate.** Platforms should provide detailed information to users about grounds for removal, with specific reference to the rule that is violated and an explanation of the possibility to request human review.

- **States should disclose all requests sent to internet intermediaries and the responses they have received,** and should mandate that platforms disclose whether any state request led to tweaks or changes to the machine-learning model they use to moderate potentially illegal content.

- **States should require that internet intermediaries grant researchers and civil society organisations access to datasets and models,** so that they can evaluate them and inform public interest-driven research. If necessary, institutional review boards and an independent accreditation process could be established.

- **States should require proof of the utility of the monitoring tools used.** For instance, they could be asked to detail use cases where illegal content was identified accurately—and where non-automated means would not have produced the same degree of success. Proof of utility is essential in addressing the necessity of the intervention. Ultimately, only proven utility can be assessed for its proportionality with the human rights harm caused.

Recommendations for user-centred transparency

- **States should ensure that internet intermediaries properly disclose that a user is or will be affected by algorithmic decision-making, including content moderation, and that users can at least opt-out of automated decision-making.** Users must be able to exercise control over content moderation detection tools, which ideally should be secured by an "opt-in" mechanism by default. Meaningful awareness enables individual users to opt in/out of automated decision-making if they wish to do so. Internet intermediaries should design consent and privacy policies in a way that facilitates informed users' choice, in line with data protection laws.

- **States should ensure that users have access to profiling data[22] that internet intermediaries hold about them, including any inferences that are made about them.** This data should be made available to users on request in a comprehensible and accessible format. Users should also be able to rectify and delete their profile. While the General Data Protection Regulation (GDPR) largely ensures this right in the European Union, there is a need for effective and accessible procedures or interfaces that allow individuals to obtain

22 The GDPR defines profiling as automated processing of data to analyse or to make predictions about individuals; meaning that "simply assessing or classifying individuals based on characteristics" could be considered profiling, with or without predictive purpose.

this information easily. Therefore, minimum standards for user-centred transparency obligations as set out in Articles 13(2)(f) and 14(2)(g) of the GDPR should be mandated by states across the OSCE region.

- **States should legally mandate internet intermediaries to provide explanations regarding the models used, input data, performance metrics and testing of their machine-learning model, in tangible, comprehensible and age-appropriate language.** Such an explanation will allow users to contest algorithmic decision-making and/or to opt out. The right to oppose the use of automated decision-making systems should apply even if a human is involved in the process.

- **States should mandate that internet intermediaries properly explain algorithmic decision-making to users.** An explanation of a particular decision should be available to users as a minimum requirement, to ensure the contestability of automated decisions in content moderation. The explanation should be in understandable language and should include statistics that were used and a detailed explanation of the intermediary's policy behind the decision.

Recommendations for transparency requirements necessary for effective access to remedy and redress for those targeted through hate speech

- **States should oblige internet intermediaries to provide reasoned decisions explaining the process and specific choices made concerning content actioned as hate speech.** The reasoned decision to act upon hate speech should be delivered to all users affected, and should be accompanied by an explanation of the rights of each concerned party and clearly formulated instructions on how to appeal the decision. The same rule should apply for counter-notices, whether they are rejected or there is a finding in favour of the content provider.

- **States should oblige internet intermediaries to preserve all data on content removals, in compliance with data protection standards.** This includes, but is not limited to, information about which takedowns did not receive human review, whether users tried

to appeal the takedown, and cases where content was reported but not acted upon. In addition, where feasible, internet intermediaries should include in their transparency reports information and statistics about the kinds of hate speech they acted upon (for instance, which protected characteristics were violated), the proportion and rate of successful appeals, and the remedies granted.

- **States should ensure that transparency requirements for internet intermediaries secure the preservation of all content classified as hate speech that is automatically blocked or removed,** including individual posts, videos, images and entire accounts. Subject to data protection and privacy requirements, this content should be made available to researchers on request, to provide additional oversight of redress mechanisms and the fairness and effectiveness of appeal mechanisms, particularly for marginalised groups.

Recommendations for transparency requirements necessary for effective public oversight

- **States should recognise and empower designated oversight bodies, with expertise in the areas of equality and non-discrimination, to monitor and address unequal or discriminatory effects of automated decision-making on marginalised groups.** These bodies might include national human rights institutes, ombudspersons, or information and privacy commissioners, and may complement the work of domestic courts. It is crucial for public authorities to enable and empower these oversight bodies to fulfil this role by providing them with adequate and meaningful legislated powers, as well as secure and sufficient resources.

- **Equality bodies should be able to undertake strategic litigation to challenge discriminatory outcomes of automated measures.** These bodies should be supported with sufficient funds and have a team of staff members that is dedicated to this particular topic and working towards enhancing transparency in the use of automated measures.

- **States should ensure that mandatory transparency reporting requirements for internet intermediaries focus on quality and not on quantity.** Figures alone only serve as a point of comparison;

they do not provide valuable information about how internet intermediaries deal with user-generated content. Therefore, internet intermediaries' transparency reports should be required to include: the number of all received notices; the type of entities that issued them, including private parties, administrative bodies, or courts; the reasons for determining the legality of content or how it infringes the internet intermediary's terms of service; and whether the content was flagged by private parties, automated tools, or trusted flaggers.

- **States should ensure that legally mandated transparency reporting provides clarity on what content moderation method was deployed:** content removal, content demonetisation, content deprioritisation, account suspension, account removal, or any other action against flagged content or users' accounts.

- **States should mandate minimum requirements for transparency reporting, including reporting:** concrete time frames for notifying the content provider before any action is taken; concrete time frames for filing a counter-notice; the exact time that will pass before the content is restricted; the timeframe for an appeal procedure; and the number of appeals received and how they were resolved.

- **Specifically in relation to hate speech, states should oblige internet intermediaries to publish the number of reports of abusive or harmful conduct they receive per year.** This should include how many of these reports are for hateful conduct targeting protected characteristics, such as race, ethnicity, religion or gender. Specific attention needs to be paid to intersectional considerations about the ways in which race, class, gender and other individual characteristics may combine into differential modes of discriminatory treatment.

- **States should oblige internet intermediaries to publish aggregated data about how many content moderators they employ per region, as well as the language in which the moderators operate.** They should provide concrete information about how moderators are trained to identify gender and other identity-based potentially harmful content, as well as how moderators are trained on international human rights standards.

- **States should provide meaningful transparency reporting on actions that they have taken in response to the spread of potentially lawful but harmful content.** Public authorities should regularly make publicly available the following comprehensive information: the number, nature, and legal basis of all content restriction requests sent to internet intermediaries; actions taken as a result of those requests; and content restrictions based on mutual legal assistance treaties.

Recommendations on data-access frameworks for independent stakeholders with relevant expertise

- **States should establish mandatory external reporting for internet intermediaries** that should be accessible to all relevant independent stakeholders and public authorities, including researchers and civil society organisations. Internet intermediaries should be obliged to enable independent external audits of any automated model, while protecting trade secrets and the privacy/security of data.

- **States should establish mandatory data access modalities and external reporting for online platforms.** Such reporting should be accessible to all relevant independent stakeholders and public authorities, including researchers, civil society organisations and affected users. Platforms should be obliged to enable independent external audits of their algorithm-driven models, while protecting trade secrets and privacy/security of data. States should establish criteria to ensure independence and competence of the auditors.

- **Any content governance legislation or policy launched by states must be evidence—and research-based.** Public authorities must be granted meaningful access to data stored by internet intermediaries, in line with adequate data protection frameworks, so that the authorities can develop evidence-based policies and ensure adequate independent public oversight. States should therefore establish data access requirements for third parties, clearly determining who can access the data, what data can be accessed, and how this data is to be gathered and vetted before disclosure, and by whom.

- **States should establish criteria to ensure the independence and competence of auditors.** Internet intermediaries should subject themselves to regular independent, comprehensive and effective audits. A description of the system's potential legal or other effects should be accessible for audit by independent bodies with the necessary competencies. Nonetheless, such risk assessments should always be a secondary measure. Ex ante human rights impact assessments conducted under public oversight should be the primary step.

- **Civil society organisations, academic researchers conducting research in the public interest and journalists should be able to conduct meaningful monitoring and audits of automated decision-making systems.** Independent stakeholders performing third-party auditing should be able to access all information they need, such as source code, data criteria and performance metrics, to conduct substantive oversight of internet intermediaries' self-regulation. The information provided should enable third parties to audit and report on the functioning, effectiveness and errors of automated decisions behind specific content takedowns—as well as content that is left on the internet intermediary.

3.2 Recommendations for respecting human rights in content governance

- **States should develop a human rights policy with emphasis on salient human rights issues, including freedom of expression, freedom of the media, privacy, non-discrimination, and right to life, liberty and security.** States are the duty-bearers under international human rights law and hold a positive obligation to protect human rights from interference by others, including private actors or individuals. They should therefore commit to adhering to international human rights law, and should ensure that national laws regulating platforms and content governance are fully compliant with the international human rights framework by which states are legally bound.

- **States should refrain from legally requiring online platforms to deploy automated tools for detection and identification of potentially illegal or harmful content,** which in some jurisdictions is referred to as a "proactive measure".

- **States should provide clear guidance on what is considered illegal content under the applicable legislative framework.** Independent judicial authorities should provide detailed assessment of what is illegal content and differentiate between various types/categories of illegal content. States should require that platforms disclose what and how different automated tools are used for specific categories of illegal content, and what are the intended goals (detection, identification, take down/removal, managing traffic access, amplifying/de-amplifying, "shadow banning", etc.).

- **States should safeguard and legally mandate human rights due diligence for algorithmic content moderation** by establishing mechanisms to mitigate adverse impacts of companies' use of AI systems for moderating and curating user-generated content, including hate speech and illegal speech. This can be achieved by obliging internet intermediaries to conduct human rights due diligence of AI systems for detecting, identifying and addressing potentially harmful content. Intermediaries should be required to assess systems' accuracy and error rates and the potential for harm of so-called false negatives and false positives, while overall working to prevent and mitigate discriminatory outcomes of AI systems, with an emphasis on freedom of expression and freedom of the media. Diverse datasets, as well as knowledge and understanding of local contexts, linguistic nuances, and coded language, are essential.

- **States should legally mandate human rights due diligence of the data-harvesting business models.** Data-harvesting business models can increase adverse human rights impacts by encouraging potentially legal but harmful content online. Internet intermediaries whose business models depend on targeted advertising and mass collection and analysis of user data should operate on an opt-in basis, where users proactively consent to data collection and personalised content moderation and curation. At a minimum, such internet intermediaries should offer the possibility to opt-out of data

collection and/or algorithmic content moderation, while providing alternative means for ensuring users' safety online.

- Safeguarding "opt-in by default" for algorithmic content moderation systems would be a desirable mechanism because it offers greater protections for users who may be less aware of how these systems operate. Internet intermediaries should design "consent" and privacy policies in a way that facilitates informed choice for users and complies with data protection laws. An "opt-in" mechanism should enable users to exercise at least some defined minimum degree of control over recommendation systems.

Recommendations for mandatory human rights impact assessment

- **States should oblige transparent, independent and inclusive ex ante human rights impact assessments (HRIAs), within a clear regulatory framework and with oversight by a regulatory agency or independent stakeholders with relevant expertise.** The assessments should include a review of intermediaries' products, services and systems and their impacts on human rights, with an emphasis on users' right to free expression and concerns related to plurality of media. HRIAs should be conducted as openly and transparently as possible, and with the active engagement of individuals and groups affected by hate speech and illegal speech. These ex ante HRIAs should be based on input from affected communities and stakeholder groups, including civil society and marginalised groups. Results of HRIAs should be made available to the public and should be accessible and easily understandable.

- **States should mandate that companies conduct human rights impact assessments** of their platforms' algorithmic content moderation models on an ongoing basis throughout the lifecycle of their AI systems. Companies should carry out meaningful engagement with external stakeholders with relevant expertise in human rights and in the design, development and deployment of systems moderating illegal content online. This engagement should emphasise inclusive and participatory approaches granted to marginalised and vulnerable groups. The set-up, methodology and results of the human rights impact assessments should follow generally recognised best practices and should be publicly accessible.

Impact assessments must also allow for addressing questions of proportionality. This in turn requires that both the utility/necessity of the intervention and the resulting human rights harm are assessed.

- **States should ensure that internet intermediaries develop internal processes that enable them to detect and prevent human rights risks.** While the structure and scale of these mechanisms will depend on the size of the intermediary, all intermediaries, regardless of their size, should establish appropriate internal mechanisms, including internal audits. Especially in the case of hate speech, the risk assessment criteria must help determine whether any individuals or groups from marginalised communities are disproportionately impacted, and if so, how. Specific attention should be paid to intersectional considerations about the ways in which race, class, gender and other individual characteristics may combine into different modes of discriminatory treatment.

3.3 Recommendations on access to effective remedy and redress

- **States should require that internet intermediaries establish operational grievance mechanisms.** First, the affected user must have the possibility to request additional information about the outcome of the algorithm-driven content moderation tool, especially if outcomes lead to removals of content. Second, the user must have the possibility to request human review. Third, users must have access to all necessary information to appeal the decision, including in judicial courts. This includes, but is not limited to, information related to the purpose of the algorithmic-driven content moderation tool, conditions of deployment, evaluation metrics (false positives/false negatives), etc.

- **In order to ensure that individuals have access to effective remedy, states should require that specific reasons are provided for content governance decisions,** regardless of whether they were taken through human or automated review. Users must be notified of content moderation decisions that concern them, including content removals, demonetisation, and account suspensions and removals.

- **States must ensure that internet intermediaries provide a meaningful opportunity for users to appeal decisions.** This is of particular importance where content is removed and accounts are suspended. Appeal processes must be accessible and timely, and provide for effective remedies, which may include restoration of removed content or overturning of account suspensions. Where initial content decisions are made by automated means, the appeal process must include human review. Clear reasons should be provided if a user-initiated appeal is unsuccessful, so that the user can understand the decision. Users should be able to provide additional evidence when appealing removal of their content or suspension of their account. The appeal procedure at the internet intermediary level can provide for remedies, such as rectification, apology, detailed reply, explanation, corrections, account restoration or combinations of several forms of remedy in one. However, this form of remedy should not replace effective judicial remedy and judicial redress. Overall, online platforms should ensure additional human review—making sure to include a "human in the loop".

- **States should encourage policies and research initiatives looking into the impact of interface design on users' behaviour, as well as tackling issues such as the deceptive interfaces known as "dark patterns".** In addition to automated detection tools, internet intermediaries continue to rely on user reports of abuse and harassment on their sites. Users have a right to appropriate redress for hate speech targeting them.

- **Internet intermediaries should improve interface design features of abuse reporting mechanisms, so that they are accessible, efficient, age-appropriate and user-centred.** Internet intermediaries should regularly gather feedback and solicit input from users and civil society, particularly those representing historically marginalised and at-risk groups, to improve the effectiveness and accessibility of reporting mechanisms. In addition, users who flag content should be advised of decisions taken and outcomes with respect to the content they reported.

- **States should ensure that community standards and terms of service that form the basis of content moderation decisions are clearly formulated and accessible.** Comprehensible rules and guidelines about permissible and impermissible uses of the internet intermediary's service, as well as the consequences of violating terms of service, must be made available. This transparency is necessary for individual users, as well as for meaningful civil society and governmental oversight. Internet intermediaries should regularly inform all users about changes in terms of service in a comprehensive and clear manner.

- **States should create a favourable environment for participation in public debate, including freedom of the media.** States should undertake preemptive and proactive efforts to address structural and institutionalised forms of hatred and the spread of hate speech online. This includes initiating and supporting awareness campaigns to educate the public—particularly users of social media platforms—about the harms inflicted on those targeted by online harassment and abuse, including societal impacts and chilling effects on marginalised groups. Such proactive efforts should also include: investments in research on the potential positive uses of AI to create safer, community-driven online spaces; initiatives to stem the tide of hate speech that go beyond takedowns or account suspensions; and creation of opportunities and fora for dialogue between internet intermediaries, civil society and marginalised groups to improve the detection and moderation of speech online.

3.4 Recommendations on the positive use of AI to create safe and community-driven spaces for marginalised groups

- **States should incentivise internet intermediaries to also give marginalised communities critical decision-making power in the process of designing and implementing new AI products.** From training data to deployment of the AI system, many people have the expertise and lived experiences to guide this process in a way that maximises positive impact and minimises externalities on historically marginalised and at-risk communities.

- **States should support existing AI initiatives by and for marginalised communities, and support efforts to empower and educate communities to understand and use potentially beneficial AI.** Some groups are already creating community guides and workshops on how to use AI to empower, rather than surveil and further marginalise communities.[23]

- **States should support a broad diversity of approaches rather than a "one-size fits all" solution.** Examples include the "Opt-out" browser add-on, which seeks to detect misogyny and remove content, like an ad-blocker. In the P2P technology space, some are advocating the use of "glasses" or subjective moderation.[24] In this case, there could be a range of AI moderation algorithms, and the individual user could choose to activate one or more systems at the same time. Such algorithmic-driven tools would not censor the entire conversation, but would simply change the content to which users are exposed individually. A diversity of strategies creates robust solutions through applied experimentation.

- **States should promote open-source of existing proprietary models wherever possible, and allow for community feedback in their implementation.** While models can be highly profitable when closed-source, projects like Hugging Face show that the AI world can be robust, profitable and open.[25]

23 For example, see https://alliedmedia.org/wp-content/uploads/2020/09/peoples-guide-ai.pdf.
24 Emmi Bevensee, The Decentralised Web of Hate: White supremacist are starting to use Peer-to-Peer technologies. Are we prepared? At <https://rebelliousdata.com/p2p/>
25 For further details, please consult https://huggingface.co/.

3.5 Recommendations to formalise cooperation with law enforcement

- **States should adequately enforce safeguards to prohibit mandatory transfer of data, especially to law enforcement,** and in this vein, take specific measures to protect marginalised and vulnerable groups.

- **When monitoring and/or tracking content online, states should adhere to international human rights law, including the three-part test for any restrictions of freedom of expression.** When ordering monitoring and tracking of content—or ordering platforms to take down content identical with or similar to content previously assessed as illegal—states shall ensure that all measures are prescribed by law, pursue a legitimate aim, are necessary and employ the least-intrusive means to effectively reach their aim. In particular, the legitimate aim of any state measure resulting in the use of AI tools to conduct content governance needs to be clearly identified, and the benefits must be explicitly shown, so that the proportionality between those proven benefits and the resulting human rights harms can be demonstrated.

4. Conclusion

The proliferation of potentially illegal and harmful content online, and the impacts of algorithmic decision-making, remain complex and nuanced issues. Nearly five decades after the signing of the Helsinki Final Act, cooperation among OSCE participating States is as necessary as ever, to tackle the new challenges of online content moderation and the spread of illegal content, as well as lawful but harmful hate speech, online. This is particularly true with the rise of powerful internet intermediaries, which act as gatekeepers and moderators of expression in this new and increasingly important (digital) marketplace of ideas.

This section of the report aims to provide a principled approach to the regulation of illegal content and lawful but harmful hate speech online, with a particular focus on the impacts of hate speech and algorithmic decision-making on marginalised groups. States are primarily responsible for respecting, promoting and implementing human rights, including freedom of expression, freedom of the media and protections against discrimination. This responsibility involves effective regulation of internet intermediaries at all stages of the process—from the design and development of algorithmic models to the remedies that must be available for affected individuals and groups.

This section of the report outlines a number of proactive, preventative and responsive recommendations, which are intended to guide OSCE participating States in this task. The recommendations address various relevant aspects, such as ensuring algorithmic transparency; undertaking appropriate human rights due diligence; providing access to effective remedy and redress; meaningfully including civil society and affected communities at all stages of the algorithmic-driven tool's lifecycle; and promoting the positive use of AI to create safe and community-driven spaces for marginalised groups.

AI in Content Curation

AI in Content Curation and Media Pluralism

This part focuses on the use of AI in content curation, addressing the impact of data-driven content recommender systems on diversity and media pluralism. This part and the next one highlighing shortcomings of AI-based content curation and targeted advertising provide human rights-centred recommendations to prevent the negative impact of AI tools in content curation on the right to freedom of opinion and expression.

1. Defining the scope of content curation's impact on media pluralism

1.1 The relevance of algorithmic content curation and data-driven recommendation systems to media pluralism and diversity

Diversity and media pluralism are core democratic principles whose quality is impacted by the rise of dominant internet intermediaries and their influence over public discourse. Internet intermediaries, in particular social media platforms, have become an important source of, access point for and key distributor of information, including news content. Information dissemination and, increasingly, aggregation occurs primarily through algorithmic content curation[26] and recommender systems. Using optimisation and analysis of human and non-human agents, these systems "deliver" personalised content customised to individual profiles, resulting in the type and amount of content to which each individual is exposed. Content recommender systems, which rank content to determine what is presented to individual users, impact individuals' freedom to seek and impart information, as well as the overall information landscape and media freedom. The design of recommender systems significantly affects what is seen online, and what remains hidden—and for whom. The

26 Content curation can be understood as a set of algorithmic and human-driven processes that support the distribution of content to audiences, such as content ranking or editorial data analysis. See at: B. Bukovska et al, Spotlight on Artificial Intelligence and Freedom of Expression #SAIFE (2020), p.19.

process of algorithmic curation is underpinned [27] by the values and goals of the algorithm's creator,[28] socio-technical factors, self-regulation (Terms of Service, for example) and state regulation. Given how ubiquitous online content has become, and the significance it has in shaping opinion and decision-making, the pivotal question arises: Where does responsibility lie in defining and implementing policy to prioritise and codify media pluralism and diversity[29] in the era of digital information?

The part of the report provides a conceptual summary of key algorithmic-curation processes and their transformative impact on media pluralism. It further provides a set of recommendations for OSCE participating States on a human-rights-centred approach to algorithmic content curation. As such, this part focuses on the impact that algorithmic content curation and data-driven recommendation systems have on media pluralism and diversity in democratic societies—and the state's role to act as the ultimate guarantor of the human right to freedom of expression and to ensure an enabling environment for this expression.

1.2 Incongruence of algorithmic content curation and freedom of expression

The ability to filter, prioritise and engage with online content based on personal preferences and interests is often at odds with the individual agency to seek, receive and impart diverse information.[30] As a basic principle, internet intermediaries typically prioritise and display content to an individual based on the system's prediction that the individual is

27 K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, The Harvard Law Review, p.1664.
28 Radsch, Courtney. "Digital Information Access." In A New Global Agenda: Priorities, Practices, and Pathways of the International Community, edited by D. Ayton-Shenker, 72–83. Rowman & Littlefield Publishers, 2018. https://books.google.com/books?id=tyJJ-DwAAQBAJ.
29 On exposure see: Philip M Napoli, "Rethinking Program Diversity Assessment: An Audience-Centered Approach" (1997) 10 Journal of Media Economics 59-74.; N. Helberger & M. Wojcieszak (2018). Exposure Diversity. In P. M. Napoli (Ed.), Mediated Communication (pp. 535-560). (Handbooks of Communication Science; Vol. 7). De Gruyter Mouton. https://doi.org/10.1515/9783110481129-029.
30 P. Leersen, The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems, European Journal of Law and Technology (2020), p.12.

likely to engage with that content. Similar to systems for personalised and behaviour-based advertisements, content recommender systems thus extensively collect data of users (and non-users) to create digital profiles, assess similarities and make inferences based on this data.

Many online platforms' business model,[31] which prioritises engagement and profit over a human rights-centred approach, can and does result in exploitative and intrusive data practices, the spread of mis/disinformation and algorithmic feedback loops.[32] It has been proven to have a negative influence on content plurality, especially regarding content created by or for marginalised communities. The model perpetuates information gaps[33] and constitutes obstacles to advocacy, thereby recreating and bolstering structural societal inequality. There is also evidence that suggests that the process of content moderation benefits those groups already dominating online spaces and narratives over marginalised groups, information and narratives.[34] Moreover, algorithm-driven content discovery (e.g. search engines) have been found to reinforce racism by suggesting discriminatory search phrases and discrepancies, particularly along racial, language and gendered lines, in depictions of members of marginalised communities.[35]

For the most part, algorithmic content curation and recommender systems are based on intermediaries' own (internal) rules, interests and assumptions, rather than democratic or public interest values.[36]

31 See: Ranking Digital Rights, It's the Business Model: How Big Tech's Profit Machine is Distorting the Public Sphere and Threatening Democracy (2020).
32 Bodó, B., Helberger, N., Eskens, S., & Möller, J, Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. Digital Journalism (2019), p.219.
33 A. Causevic and A. Sengupta, Whose Knowledge Is Online? Practices of Epistemic Justice for a Digital New Deal, IT for Change (2020), Retrieved from: https://itforchange.net/digital-new-deal/2020/10/30/whose-knowledge-is-online-practices-of-epistemic-justice-for-a-digital-new-deal/.
34 B. Marshall, Algorithmic misogynoir in content moderation practice, Heinrich-Böll-Stiftung (2021), p.7,11. See also: M. E, Mazzoli and D. Tambini, Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020), p. 44.
35 Safiya Umoja Noble, Algorithms of Oppression How Search Engines Reinforce Racism, NYU Press (2018).
36 C. Radsch. "Digital Information Access." In A New Global Agenda: Priorities, Practices, and Pathways of the International Community, edited by D. Ayton-Shenker, 72–83. Rowman & Littlefield Publishers, 2018. https://books.google.com/books?id=tyJJDwAAQBAJ.

Content recommendation is crucial for the growth and dominance of large internet intermediaries, and it lies at the heart of their business models. As recommender systems are "a key logic governing the flows of information on which we depend",[37] internet intermediaries are enabled to act as gatekeepers of information and knowledge. This has broader implications for public interest, representation and power (in)equality, on- and offline.[38] Intermediaries' recommender systems have significantly reconfigured the logic of public communication, including access to news, critical information and overall content in the public interest. Thus, their recommender systems significantly restrict equal access to, and of, journalists and media outlets, while pressuring professional journalism due to the outflow of advertising money to intermediaries. Recent research findings on algorithmic prioritisation, defined as "the range of design and algorithmic decisions that result in prominence and discoverability of content" [39]reveal the potential for the polarisation of opinions and attitudes online. For instance, an important factor in content prioritisation processes is individual political predisposition and/or affiliation. Prioritisation can, therefore, reinforce and perpetuate polarisation of opinions and attitudes online, especially among those users at the edges of the political spectrum who likely already consume a predominance of affiliated content.[40] It has also been shown that "some groups in society are more prone to selective exposure than others".[41]

37 T. Gillespie (2018). Custodians of the internet. Retrieved from https://www.researchgate.net/publication/327186182_Custodians_of_the_internet_Platforms_content_moderation_a nd_the_hidden_decisions_that_shape_social_media

38 P. Leerssen, The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems. Retrieved from file:///Users/eliskapirkova/Downloads/Leerssen%20EJLT_corr.pdf.

39 M.E. Mazzoli and D. Tambini. Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020), p.12.

40 B. Stark, D. Stegmann, Are Algorithms a Threat to Democracy? The Rise of Intermediaries: A Challenge for Public Discourse. Retrieved from https://algorithmwatch.org/wp-content/uploads/2020/05/Governing-Platforms-communications-study-Stark-May-2020 -AlgorithmWatch.pdf

41 B. Bodó, N. Helberger, S. Eskens & J. Möller, Interested in diversity: The role of user attitudes, algorithmic feedback loops, and policy in news personalization. Digital Journalism (2019), p.15.

While active personalisation based on user input tends to produce a greater diversity of information, passive personalisation based on algorithmic content selection tends to exacerbate the so-called filter bubble effect.[42]

Bias and discrimination, including gender-based discrimination, in data-supported algorithmic decision-making can occur for several reasons and at many levels in content curation systems, and they can be difficult to detect and mitigate. It has been suggested that the exclusion of sensitive/identity-based information sufficiently protects against discrimination. Yet discrimination can and does occur, despite these "protections", given the expansive and diverse information contained in algorithm-informing datasets. Bias in algorithms can stem from design and implementation, including unrepresentative or incomplete training data, or reliance on individual, experiential, or values-informed data that reflects historical/structural inequalities. Algorithmic bias can have a collective, disparate impact on communities, especially marginalised groups, even when there is no intention to discriminate. An exploration of both intended and unintended consequences of algorithms is thus necessary. Current public policies may not be sufficient to identify, mitigate, and remedy the impact on individuals or society at large. In addition to deliberate efforts to shape individual attention (direct manipulation), there is also a danger of unwanted and indirect biases being introduced into the algorithm through incorporation of big data at various levels of the content curation. Both direct and indirect discrimination caused by algorithms using big data are among the most pressing dangers of algorithm-driven curation processes.

The combined effect of content filtering and personalisation practically creates layers of restriction in terms of discoverability, and thus accessibility, of diverse media content. The aforementioned issues have serious implications for media pluralism, understood as a plurality of information sources (external pluralism), and of content (internal pluralism).[43] More specifically, and in the context of this report, media

42 D. Wagner, Artificial Intelligence and Disinformation as a Multilateral Policy Challenge https://www.osce.org/files/f/documents/d/0/506702.pdf.

43 On exposure see: P. M. Napoli, "Rethinking Program Diversity Assessment: An Audience-Centered Approach" (1997) 10 Journal of Media Economics 59-74.; Helberger, N., & Wojcieszak, M. (2018). Exposure Diversity. In P. M. Napoli (Ed.), Mediated Communication (pp. 535-560). (Handbooks of Communication Science; Vol. 7). De Gruyter Mouton. https://doi.org/10.1515/9783110481129-029.

pluralism also refers to the distribution of communicative power (or "voice") in society. A fair distribution of "voice", as a precondition, requires the deconcentration of power and decentralisation of resources within the information ecosystem,[44] as well as support for alternative models that offer a diversity of narratives and content. It is clear that algorithmically driven content curation processes are transforming the notions of media pluralism and diversity, which are necessary for democratic, public debate and inclusive societies.

It is against this background that state and non-state actors, primarily internet intermediaries and media organisations, but also international and regional organisations, civil society representatives and academia, are called upon to adopt policies that contribute to an enabling environment for media plurality. This means enabling access, availability, discoverability and consumption of different kinds of (media) content through different mediums and via multiple channels.

2. Algorithmic content curation and data-driven recommendation systems: impact on media pluralism

2.1 Typology

A distinct source of influence, and thus communicative power, of internet intermediaries and social media companies lies in their content recommender systems, which also "lend gravitas to their role in democratic culture".[45] In essence, a "recommender system" includes various technologies that filter, retrieve and organise information for individuals. The factors for ranking can include the level of engagement with the specific content, the type of content, when it was first shared, or how users have interacted previously with similar content. By ranking content, these systems have the potential to shape and impact individuals' ability to form opinions.

44 M. Moore and D. Tambini (eds) (2018) Digital Dominance: The Power of Google, Amazon, Facebook, and Apple. New York: Oxford University Press.
45 K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, The Harvard Law Review, p.1663.

The main purpose of content recommenders is to filter large amounts of information online. This algorithmically driven process works in different ways:

- **Content-based filtering**: individuals get content recommendations based on their stated or implied preferences. For example, if someone likes classical music or news about a favourite sports team, then the recommender system will prioritise these items that align with their interests, and will, likely, encourage engagement.

- **Collaborative filtering**: individuals get content recommendations based on people with whom they are closely associated or with whom they share similarities (in demographic category, content preferences, etc.). For example, when reading news, a system recommends articles a friend has shared/read, or, when doing online shopping, the system recommends items that people with a similar shopping history have purchased.

- **Hybrid filtering**: a combination of the above-mentioned filtering and curation methods. For example, recommending a news article a friend has liked, but only if it covers a certain topic of perceived interest to the user, and combining it with a wide range of different metadata, such as an individual's location, usage history, etc.

All of these processes are based on users' data, profiles and interactions with a given platform, as well as information gleaned from the underlying ad tech architecture. The algorithm specifies the precise way in which content recommendations are generated, using content-based and collaborative filtering. The system creates a recommender strategy for how data is combined to calculate potential engagement, based on user recommendations, to satisfy optimisation criteria. Put simply, algorithmic content curation is the strategy used by a recommender system to determine how collected data can best be utilised to reach pre-defined optimisation goals.[46]

46 To reach the optimisation goal, the algorithm can emphasise different ways of how to prioritise the collected data. For example, an algorithm could favour recency of a news article. Another strategy would be to look at popularity of articles as a ranking criterion of how to sort the final recommendations calculated for each user. Accuracy is another frequently used way of curating content. An accuracy-optimised approach tries to model user preference as closely as possible. They calculate recommendations that fall in line with existing user preferences. Depending on the data that was collected and the nature of the item in question, there are multiple ways of how to refine each strategy.

All large internet intermediaries, and social media platforms in particular, use so-called "open content recommender systems".[47] These systems utilise user-generated content in the recommender's source pool by default, but certain content items can be excluded based on, for instance, a violation of Terms of Service. To optimise engagement, these systems "personalise" the online experience by prioritising content that is assumed to be appealing and is generated through each individual's prior engagement and behaviour. Accordingly, the videos, search results, news articles or any other type of content that is displayed to the user is unique to their experience and differs from what other users see. For this reason, among others, algorithmic content recommender systems have the potential to undermine and disrupt democratic processes.[48] They risk narrowing individuals' exposure and access to different points of view, values and narratives, thereby threatening pluralism and diversity. This may necessitate state intervention and attention.

The algorithmic filtering and adaption of online content based on speculated personal preferences and interests decreases the exposure to a diversity of information, with potential negative effects on diversity and public discourse, as well as privacy. Content curation systems can therefore profoundly influence the information sources that form the basis for arriving at well-informed opinions, and thus the thought process of individuals. While the research is not yet conclusive, this could undermine individuals' ability to form their opinions and make them vulnerable to manipulative interference. As the systems are built on intrusive data practices and persuasion architectures (at scale), the inevitable personalisation of content might have a significant effect on the cognitive autonomy of individuals and interfere with their right to form an opinion.

47 As opposed to a closed recommender system that provides items to users from a limited list of options. These lists are curated by the platform owner.
48 N. Helberger (2019) On the Democratic Role of News Recommenders. Digital Journalism 7(8). Routledge: 993–1012. DOI: 10.1080/21670811.2019.1623700.

2.2 Curation and prioritisation of public interest content

The methods by which online platforms curate content through recommender systems are not transparent, and they are very rarely subject to public and/or state scrutiny. When internet intermediaries incorporate diversity into recommender systems, it is typically a design choice to engage users and increase profits. Platform curated diversity has primarily been utilised to optimise financial gain, rather than to promote democratic debate, through a practice of prolonged engagement to achieve what is referred to as an optimisation goal—increased ad revenue or a higher platform/service valuation through increased traffic.[49] Put simply, business-driven content curation benchmarks have largely been established to optimise economic gain and leverage user engagement for corporate interest, rather than seeking to reflect and ensure genuinely diverse content.[50]

Content recommender systems may also have unintended consequences from the perspective of broader societal objectives and can negatively shape and interfere with the absolute right to freedom of thought and opinion.

In addition, the processes of internet intermediaries' recommender systems typically exclude individual users' choice, control and agency—prerequisites to ensuring individual autonomy in seeking and imparting a variety of information and ideas. Following the public disclosure of a number of vulnerabilities of recommender systems,[51] there has been increasing public and state pressure to ensure that their processes better and more meaningfully prioritise "diversified" media exposure. In particular, concerns have been raised in the context of the legality and reach of political speech, and the spread and normalisation of specific value systems as protected speech, even when content is in violation of Terms of Services or international human rights standards. Given the complete lack of information on the way in which platforms govern and prioritise

49 In the words of a Facebook official: "Facebook is profitable only because when you add up a lot of tiny interactions worth nothing, it is suddenly worth billions of dollars.", K. Klonick, New Governor, p.1627.
50 K. Klonick, The New Governors, p.1664.
51 A well-known and most cited case was that of the "Napalm Girl", a journalistic photo which was taken down by Facebook based on its nudity policies.

speech, it is clear that recommender systems and correlated logics of optimisation could undermine the "fair opportunity to participate" [52] for all. At the same time, it should be recognised that human rights-based recommender systems can positively affect pluralism, for example in the contexts of authoritarianism and media capture.

It is important to define what constitutes public interest with regards to algorithmic content curation, and how it affects prioritisation of different types of content. The concept and definition of public interest content is as highly contested as the definition of diversified exposure. In principle, public interest content constitutes that information that "the public would have an interest in being informed about".[53] Another way to think about public interest content is as content that is relevant to the well-being of citizens, the life of the community or the local population. Obvious examples include COVID-19 pandemic information or information related to democratic voting processes. The volume of related, and not always reliable, content has prompted intermediaries to (publicly) prioritise accuracy. Several platforms have, in a comparatively short amount of time, demonstrated their capacity to re-configure algorithm recommender systems in an effort to filter out, or label false information, and prioritise content from trusted public health authorities. The success of these efforts, or logic behind the motivation for these changes, however, remains hotly debated,[54] while a lack of transparency into the underlying data and content moderation choices made by the platforms remain a mystery. Debates aside, internet intermediaries and social media platforms—facing increasing demands from the public and states that they be held responsible for public health awareness—have shown their capacity for reflection, and for restructuring how they prioritise and rank content.[55]

52 K. Klonick, The New Governors, p.1664.
53 M.E. Mazzoli and D. Tambini, Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020), p. 13.
54 M. Cinelli, The COVID-19 social media infodemic, The Nature (2020), p.10; See also: Global Disinformation Index, Why is tech not defunding COVID-19 disinfo sites? (2020), Retrieved from: https://disinformationindex.org/2020/05/why-is-tech-not-defunding-covid-19-disinfo-sites.
55 European Commission, Joint communication to the European Parliament, the European Council, The Council, The European Economic and Social Committee and the Committee of the Regions, Tackling COVID-19 disinformation - Getting the facts right, JOIN(2020) 8 final, 10 June 2020., section 5.

This points to a need for greater public attention and political pressure on platforms to make transparent their recommender processes and restructure recommender systems and their optimisation goals, in order to address structural problems of our contemporary media environment. The issue goes beyond questions of content governance, and additionally concerns competition law, media ownership and concentration rules.[56] It also highlights the urgent need to prioritise media pluralism and diversity policy objectives and interventions for a more enabling digital space.

2.3 News aggregation and media plurality

News aggregators function as a central hub of online news distribution, directing readers to news items, and other content deemed (by the news aggregator) to be news. This process is predominantly carried out by algorithms, which is why news aggregators are sometimes referred to as "algorithmic gatekeepers".[57]

News aggregators often involve a tension between "algorithmic logic" and "editorial logic".[58] "Algorithmic logic" significantly impacts diversity as well as political discourse by prioritising novelty, for example, over other criteria of newsworthiness (e.g. public relevance, diversity, etc.). A study of content curation processes behind AppleNews, which employs both human moderation (in the Top Stories) and algorithmic content curation (in Trending Stories), showed that human moderated content featured "more diverse and more equitable source distribution than algorithmically-selected" stories.[59] Trending Stories, according to the same study, almost exclusively included "soft news" (e.g. stories about celebrities), with Top Stories reserved for "hard news" (e.g. political content).[60] These practices were found to severely impact source plurality, and news distribution, and therefore content plurality, in two ways: First,

56 M. E. Mazzoli and D. Tambini, Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020), p. 23.
57 Napoli 2014.
58 T. Gillespie, PJ Boczkowski, KA Foot, Media technologies: Essays on communication, materiality, and society, MIT Press (2014).
59 J. Bandy and N. Diakopoulos, Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News, Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020), p.43.
60 Ibid.

aggregators created a so-called "market-expansion effect" because they provided individuals exposure to news outlets with lower popularity or brand awareness. Second, the deployment of aggregators incentivised some users to limit or stop direct use of news outlets, resulting in the so-called "substitution effect". Since user reactions are usually based on first impressions, clickbait is used in the news feed to attract attention and engage users, thereby facilitating advertising that generates profit. This approach further challenges media sustainability, and consequently independence and pluralism, adding to the overall pressure and financial constraints faced by legacy media because of internet intermediaries' concentrated advertising and data exploitation models.

There is a growing imbalance between the outreach and communications impact of legacy media and online platforms, and content creation vs. content curation. Given that the traditional subscription-based business model is in decline, legacy media are struggling for viability. An ever-growing number of people get their news exclusively from other sources, where articles are more likely to be available "for free". The willingness to pay for quality news has decreased, while usage of "free" news aggregator sites and social media platforms increased. As a result, many online news outlets have no choice but to search out new revenue streams. They are coerced into adopting many of the practices used by the large platforms, which exert significant power over the logic of the digital adtech industry (e.g., by employing targeted advertising, publishing sponsored content, or collecting and selling user data). This trend has profoundly negative implications for media freedom globally: It creates an environment in which legacy media organisations must compete with social media companies and intermediaries for the same revenue sources, while also being subject to intermediaries' recommender systems and content curation policies. The situation contributes to an erosion of trust in media, decreased responsibility for creations and dissemination of disinformation and other problematic content.

Some legacy media organisations also employ algorithm-driven tools themselves, with content personalisation and optimisation playing integral roles in media production processes. There remain stark differences between "news logic of personalisation"[61] and "platform logic of personalisation", with news media subject to a systemic lack of technological and financial resources, devaluation of traditional editorial and professional ethics, the prevalence of newly emerging private interests, and economic incentives. Algorithmic content curation models and recommender strategies developed and deployed by independent legacy media outlets, and especially public service broadcasters, could offer alternative models for better ensuring audiences' exposure to diversity, potentially even offering individuals with models for "diversity by design".[62]

Evidence shows that journalists value "editorial logic", such as "transparency, diversity, editorial autonomy, broad information offer, personal relevance, usability, and surprise" over the business-driven algorithmic logic of recommender systems.[63]

Algorithm-driven content curation and recommendation processes and practices pose threats to media plurality and diversity and raise concerns regarding the full enjoyment of the right to freedom of expression. The major contributing factors constituting these threats are:

- Financial instability of, and fiscal pressure on, legacy media: Online platforms have gained enormous economic power, primarily through advertising revenue, and they use this leverage to dictate conditions

61 B. Bodó, Selling News to Audiences – A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media, Digital Journalism (2019), p.17-18.
62 See more about this content in N.Helberger, Diversity by design—Diversity of content in the digital age, Government of Canada (2020), p.8; Natali Helberger, Kari Karppinen & Lucia D'Acunto (2018) Exposure diversity as a design principle for recommender systems, Information, Communication & Society, 21:2, 191-207, DOI: 10.1080/1369118X.2016.1271900.
63 This study involved newsrooms from the Netherlands and Switzerland; M. Bastian, N.Helberger & M. Makhortykh, Safeguarding the Journalistic DNA: Attitudes towards the Role of Professional Values in Algorithmic News Recommender Designs, Digital Journalism (2021), p.21.

for the curation of all online content, including editorial media and news content. This power imbalance includes an imbalance of "opinion power"[64] and the power to "influence processes of individual and public opinion formation", which in turn enables "these platforms [to] change the very structure and balance of the media market, and thereby directly and permanently impact the pluralistic public sphere."[65]

- Legacy media, compelled to adopt similar business models and social media "logic", miss an opportunity to change the "rules of the new communication orders",[66] and to contribute to a more diversified media landscape. Yet there are alternative algorithmic-curation models centring on public interest content and professional journalism practices—typically instated by legacy and public service media organisations—and they do offer alternative models to mitigate the potential problems caused by a lack of prioritisation of public interest content.[67]

- Improving algorithmic content curation with a goal of increasing diversity of media sources poses several challenges:

    - Even if internet intermediaries and social media companies "train and game" algorithms "for good"—to expose heterogeneous audiences to heterogeneous content—these practices lack meaningful transparency, and individuals have no agency regarding the design and logic that govern these systems. This presents a significant and systemic risk to the enjoyment of freedom of expression.

    - While personalised content and optimisation processes could contribute to meeting diverse individual, group and societal needs—and generate potential for diversity—such potential should be driven by public policy objectives and corresponding interventions.

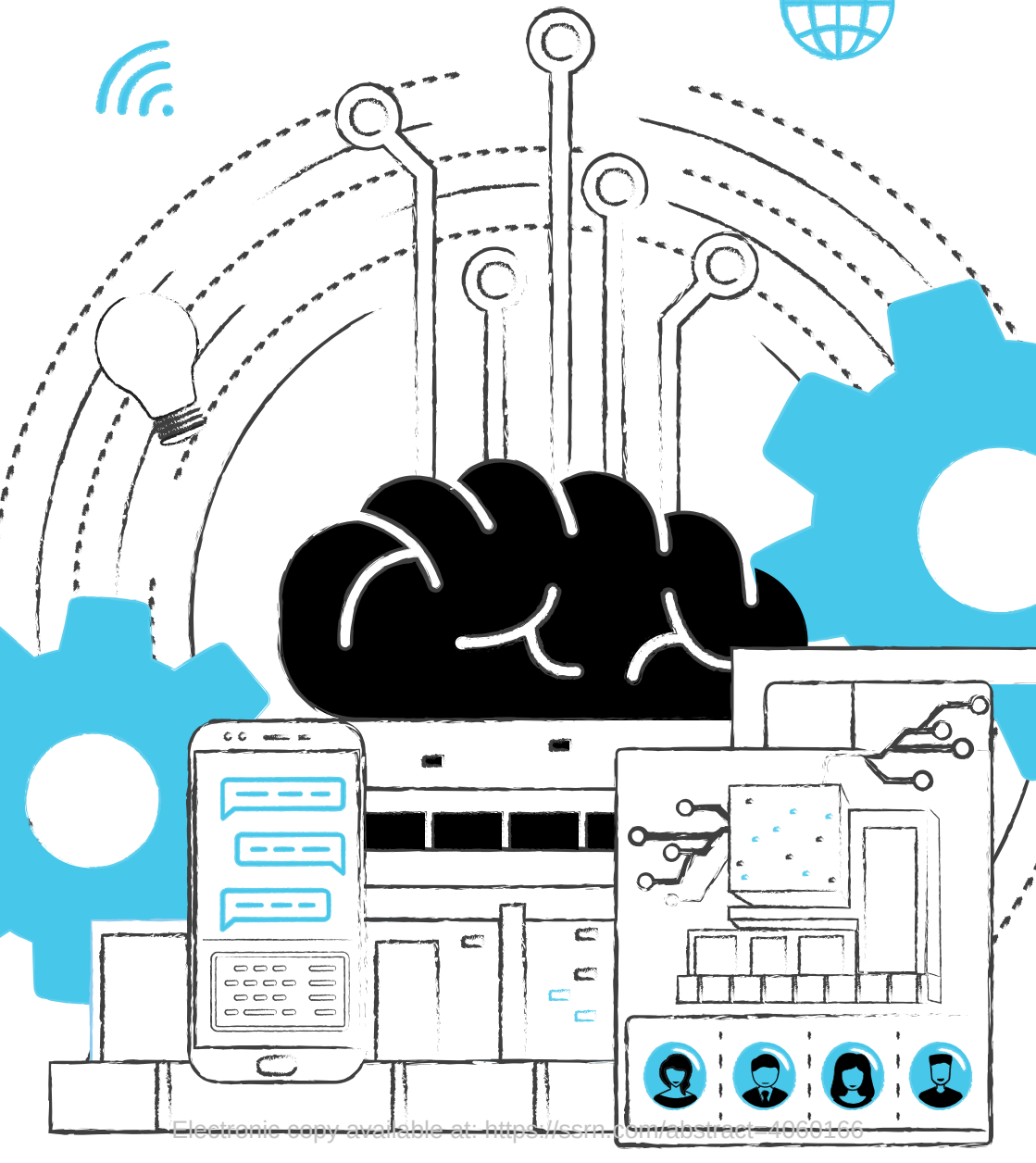
64 N. Helberger, The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power. Digital Journalism, 8(6), 842-854 (2020)
65 Ibid., p.846.
66 For an in-depth discussion about this problem, see: N. Helberger,The Political Power of Platforms: How Current Attempts to Regulate Misinformation Amplify Opinion Power. Digital Journalism, 8(6), 842-854 (2020)
67 M.E,Mazzoli and D.Tambini,Prioritisation uncovered: The Discoverability of Public Interest Content Online. Council of Europe (2020).

- There is little to no information on how content produced by and for marginalised communities circulates online, and how recommenders treat such content. Studies suggest[68] that certain content and speech is treated differently, giving rise to concerns that content is not equally accessible, and that safeguards to prevent discriminatory algorithmic outcomes and ensure fair and equal public participation and deliberation have not been developed or implemented.

- Consequences of algorithm-driven content curation can compound human rights abuses and rule of law violations when amplified in certain national contexts, specifically in conjunction with systemic (state-led and private) media capture and monopolised control of public dialogue. Under these circumstances, additional layers of algorithm-driven restrictions to and for media pluralism and diversity intensify the aggregate sum of individual loss of the right to freedom of expression.

68 See, for example: A. Chinmayi, Facebook's Faces, Forthcoming Harvard Law Review Forum Volume 135 (2021) and K. Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, The Harvard Law Review (2018); C. O'Neil, Facebook's VIP "Whitelist" Reveals Two Big Problems, Bloomberg Opinion (2021), Retrieved from:https://www.bloomberg.com/opinion/articles/2021-09-15/facebook-s-xcheck-vip-whitelist-reveals-two-big-problems

Media Pluralism



3. Human rights-centred recommendations on the use of AI in content curation

States are primarily the guarantor of media pluralism under the international human rights protection framework. They are to act as the ultimate guarantors for the enjoyment of human rights, including responsibility for an enabling environment for the rights to freedom of expression and freedom of the media. The following recommendations for OSCE participating States, generated during the workshop, focus on: strengthening a pluralistic media landscape and plurality of voices (3.1); fostering an enabling environment for the diversity of media content and individual exposure to diverse media (3.2); and enabling individual agency and control (3.3).

3.1 Recommendations on strengthening a pluralistic media landscape and the plurality of voices

This part of the report seeks to offer participating States a normative agenda that fosters a pluralistic media environment and "the coexistence of diverse and competing interests—that is a basis for a democratic equilibrium."[69] This agenda is constrained by the shrinking opportunities for a democratic-driven media space, the unbalanced digital platform dominance and excessive market concentration. Participating States should ensure conditions for media innovation, independence and sustainability, especially public interest-driven media, and enforce models of content creation, curation and distribution that foster these conditions.

- **States should ensure, through regulatory initiatives, a level playing field for all media actors, by removing obstacles for the provision of fair and effective market conditions.** The resulting market conditions should enable all media to access and use new technologies and to develop alternative business models—including alternative models of algorithmic content curation that foster a diversified media landscape and the proliferation of public interest content.

69 A. Roksa-Zubcevic et al, Media Regulatory Authorities and media pluralism, Regional Publication. Council of Europe (2021), p.12-14.

- **States should analyse how existing and future media pluralism-related policy addresses the issue of public interest content,** especially in light of the significance of online platforms in distributing public health information during the COVID-19 pandemic.

- **Public-private partnerships between states and social media companies and other intermediaries should be rigorously transparent** and subject to citizen oversight and public scrutiny. This should include the regulatory framework of the media pluralism landscape.

- **States should employ policy and legislation to prevent the unbalanced and monopolised market power that currently exists,** especially with regards to internet intermediaries and state-controlled content distribution. Any and all state intervention must ensure a pro-democratic regime that is genuinely independent and offers structural solutions to bolster plurality.

- **States should promote plurality and technological and media innovation** by funding holistic independent research that helps media actors, public oversight institutions and academia understand the current distribution of power—especially regarding the effects of recommender systems, analysis of recommender logic, and the resulting impact on media pluralism and diversity.

3.2. Recommendations on fostering an enabling environment for diversity of media content and individual exposure to pluralistic information

This part of the report addresses whether and how internet intermediaries should ensure individuals' equal access to, and participation in, public spaces, by examining diversity as a normative concept.

- **States regulatory and policy interventions should preserve and foster the internet as a space for democratic participation and representation.** Any state regulation of the digital space should

have a clearly defined scope that is necessary for, and proportionate to, a transparent objective, in full compliance with the international human rights framework.

- **States should engage in and support cross-sectoral dialogue to gather the most current and relevant data on the impacts of algorithmic content curation,** such as polarisation, informational gaps, etc. Independent oversight and transparency of diversity monitoring requires **a multidisciplinary approach, led by academic institutions or civil society organisations, with support from the state.** Intersectional diversity monitoring should be used to identify content and audiences that are at risk of, or have historically faced, exclusion from public participation and/or representation.

- **States should adopt an inclusive approach and ensure multi-stakeholder participation and ownership of algorithmic content curation.** Democracies are not self-perpetuating systems. For democracies to thrive, citizens must have the ability to make informed decisions. Through the provision of open dialogue and cross-sectoral collaboration with internet intermediaries, states—together with civil society organisations, marginalised communities, media organisations, journalists and their representatives—could foster sustainable, cross-sectoral cooperation, including between state and non-state actors. Moreover, states should push for diversity in the teams of developers who create algorithmic content curation systems, so that diverse interests and perspectives are represented in the design and implementation of the algorithms.

- **States should provide support and resources to existing independent media regulatory bodies** that use a process of co-creation and inclusion of all national media actors and experts to support an economic, legal and political environment in which diversity is cultivated as a core democratic objective.

- **States should develop an evidence- and research-based legislative framework to ensure accountability of internet intermediaries, including by mandating human rights due diligence.** Human rights impact assessments should be part of any risk mitigation strategy or any external audits to ensure public oversight.

- **States should strengthen independent media regulatory bodies, and other competent institutions, and involve them in public oversight and research.** For instance, these bodies should be involved in human rights impact assessments to address risks that internet intermediaries pose to diversity and plurality, including risks posed to marginalised communities. Such assessments should be accompanied by accountability mechanisms, and there should be transparent disclosure and publication of assessments, audits, and the like.

- **States should increase public funding for independent, quality journalism and/or provide financial resources to independent stakeholders** with relevant expertise and a proven human rights record. These independent actors can offer alternatives to the existing revenue-oriented and data-driven business models, thus fostering decentralised technological algorithmic curation systems that promote public values, such as media diversity, inclusivity and tolerance.

- **States should ensure that any potential intervention in this field does not limit the positive functionality of personalisation or of media independence,** while at the same time providing support and intervention to ensure that content diversity and public interest content is the design-focus. Personalisation can be valuable to individuals when it is used to refine searches and speed up the retrieval of information.

- **States should establish and safeguard adequate data access frameworks that enable vetted researchers, civil society organisations and other independent stakeholders, such as the media, to access data held by internet intermediaries.** At the same time, abuse of such a framework needs to be prevented through the use of ethics guidelines or the creation of an independent authority with an overseer function.

3.3. Recommendation on enabling individual agency and control

The empowerment of individuals should be built-in to algorithmic design, and similarly, public interest and human rights-centred design should be at the forefront of algorithmic deployment.

- **States should support self- and co-regulatory initiatives, and create conditions that codify individual control over what is seen online.** This could be achieved by legally mandating options such as opt-in by default for content recommender systems, and easy identification and choice for defining editorial and non-editorial personalisation.

- **States should mandate transparency and explainability of pre-selected news personalisation and data processing.** This should include transparency of the criteria, principles and types of arrangements driving content prioritisation decisions (to foster public trust, and to allow the public to understand whether commercial or public interest objectives are considered). Participating States should also require intermediaries to have due process measures in place. For example, when intermediaries make restrictions to news feeds they should inform the individual affected about their respective policies and provide effective redress mechanisms. Similarly, **states should support self- and co-regulatory initiatives that ensure intermediary transparency** on the processes driving content prioritisation decisions.

- **States should establish sustainable media and digital literacy programs for all societal groups**. Individuals are often not aware of, and/or do not realise, the implications of algorithmic content curation for their enjoyment of human rights and fundamental freedoms.

- **States should put a special focus on fostering the right to access, seek and impart opinions and ideas of all kinds among all age groups. In particular, states should empower the process of individual opinion formation,** including for young people, who are regularly deprived of proper access to legacy media content.

4. Conclusion

Content curation has, at least from the context of media pluralism and media diversity, largely been pushed to the margins of broader content governance discussions. This omission, compounded by a lack of understanding of the importance of media and informational diversity for heterogeneous audiences, has implications for harm that are equally as concerning as those risks stemming from illegal content and dis/misinformation. All algorithm-driven processes of content curation and moderation are intrinsically connected and must be addressed as such.[70] These processes are particularly important as algorithms determine what individuals see, what information is prioritised and what content is excluded. Online gatekeepers increasingly rely on recommender systems that systematically analyse patterns of user behaviour and create profiles to determine what information is more likely to engage a given user. In other words, gatekeepers harvest data to determine what personalised content to offer to individual users, in order to spur their engagement and generate more data about them—even if the decisions made are at odds with democratic discourse, the diversity of information, media pluralism and the right to privacy. For this reason, and as clearly articulated in the recommendations, inter-disciplinary research, and transparency of intermediary policies and content curation practices, are crucial preconditions for centring media pluralism and diversity in algorithm design and deployment.

This outcome report highlights the problematic nature of personalised content recommender systems used by internet intermediaries and social media platforms in particular. It outlines the concerning implications these systems have for societal cohesion, diversity, the quality of information within the public discourse and privacy. At the individual level, the online experience is strategically coloured by decisions made for profit, implemented through algorithms without the awareness of affected individuals or the scrutiny of public authorities, and based on intrusive data collection and analysis that are designed to circumvent privacy and data protection laws—with largely negative effects on the diversity of information, media pluralism and the right to privacy.

70 In simple words, over-removal of "legitimate" content is in fact also a risk for media diversity.

There exists ample evidence that online platforms' opinion power has the ability to steer and amplify certain public narratives and types of discourse over others. For countries with fragile or oppressive political systems, this opinion power, coupled with algorithmic amplification, can have disastrous consequences for individual enjoyment of human rights. By ranking and differentiating content and recommendation outcomes, internet intermediaries are reconfiguring the public debate in a way that empowers those already in privileged positions. The price for this is a reduction in diversity in general, and particularly disadvantageous for historically marginalised groups, who continue to be pushed to the margins of public discussions in a process that recreates and bolsters inequality and injustice. While algorithmic content curation has the power to limit participation, create division and limit the spread of information, media diversity fosters social cohesion, tolerance and distribution of communication power. It is up to states, primarily, and also non-state actors, most notably intermediaries and media organisations, to ensure that media plurality, equal access and the full enjoyment of human rights are the basis for rules affecting the online information space, as these are building blocks for truly democratic digital societies.

AI in Content Curation and Surveillance-Based Advertising

This part focuses on the use of AI in content curation focusing on the nexus between surveillance capitalism and targeted advertising, and the resulting impact on freedom of opinion and expression. It highlights shortcomings of AI-based content curation and targeted advertising and provides human-rights-centered recommendations for OSCE participating States to address the negative impact that AI tools in content curation have on the right to freedom of opinion and expression.

1. Defining the scope of the impact of surveillance-based business models in their use for content curation

1.1 Impact of automated decision making on the right to freedom of opinion

The international human rights framework distinguishes between the internal and external dimension of the right to freedom of opinion. While the external dimension of this right can be subject to legitimate, proportionate, and non-discriminatory restrictions that are necessary in a democratic society, the internal dimension of the freedom of opinion, so-called forum internum, is absolute and non-derogable.[71] Article 19 of the Universal Declaration of Human Rights as well as the International Covenant on Civil and Political Rights protect this absolute right from any restriction or interference. In the words of the UN Special Rapporteur on Freedom of Expression and Opinion, "any involuntary disclosure of opinions is prohibited and mental autonomy is affirmed."[72]

The data-harvesting business models of large online platforms enable the

71 Office and the High Commissioner for Human Rights, CCPR General Comment No. 22: Article 188 (Freedom of Thought, Conscience and Religion), available at https://www.refworld.org/docid/453883fb22.html, 1993.
72 Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan, Disinformation and freedom of opinion and expression, available at: https://undocs.org/A/HRC/47/25, 2021.

advertising industry to develop or rely on data-driven targeting strategies. Through this approach, companies identify and exploit people's or communities' behavioural patterns and characteristics. The umbrella term that covers these manipulative techniques is "surveillance-based advertising", understood as a blanket term for digital advertising that is targeted to individuals or groups, usually through tracking and profiling based on personal data. The context of where a specific ad is placed can be random, because, as it is targeted at individuals, it can follow them around in different contexts.[73] In most cases, surveillance-based advertising is part of an automated process, by which each individual ad is chosen and placed in a matter of milliseconds. This means that neither the ad publisher (e.g. the owner of a website or app) nor the advertiser (e.g. the owner of the brand that is promoted) chooses which ads to show to whom, or where to display them. This is automatically decided by technological systems that are often controlled by third party intermediaries (so-called "adtech" companies).[74]

Surveillance-based advertisement has significantly contributed to the exploitation of people's particular characteristics to increase the persuasiveness of a message, thereby unjustifiably interfering with their absolute freedom to form an opinion and to enjoy independent thought processes. People who are using platforms' services are being manipulated to think or to make decisions they would have otherwise perhaps never made. Surveillance-based advertisement exploits individuals' vulnerabilities even if it does not directly identify those vulnerabilities. Through the use of so-called "lookalike audiences", advertisers can duplicate people's groups with certain characteristics in order to reach new individuals that share the same characteristics. Automated tools and the dominance of a few online platforms has enabled greater manipulation as every single individual using their service can be targeted all the time and at any time.

73 Norwegian Consumer Council, Time to ban surveillance-based advertising: The case against commercial surveillance online, available at: https://www.forbrukerradet.no/wp-content/uploads/2021/06/20210622-final-report-time-to-ban-surveillance-based-advertising.pdf, 2021.

74 Norwegian Consumer Council, Out of control: How consumers are exploited by the online advertising industry, available at: https://fil.forbrukerradet.no/wp-content/uploads/2020/01/2020-01-14-out-of-control-final-version.pdf, 2020.

There is increasing demand for a ban on practices that adversely impact people's absolute right to freedom of opinion and freedom of thought, in particular as individuals' thoughts and opinions are being widely influenced without their knowledge or consent. This phenomenon specifically includes targeted behavioural tracking and individual cross-site/cross-device tracking. Such constant invasive corporate surveillance poses a risk of systematic manipulation of individuals, beyond traditional forms of advertising influence. Surveillance-based advertising targets individuals in opaque ways[75] and may exploit vulnerabilities, opening new possibilities for manipulation. In particular, when combined with revenue-maximising algorithms for content curation, surveillance-based advertising may impact how individuals speak out and behave, effecting diversity of information, views and opinions.

Despite online platforms' claims that there is no turning back from surveillance-based advertising, the internet was not built on a "creepy ad" business model. In fact, quite the opposite. States must avoid directly or indirectly protecting business models that stand on surveillance-based advertisement and violate international human rights law. Ending abusive models also means opening the door to human rights compliant alternatives, including innovative forms of contextual advertising that rely on minimum personalisation and no individual targeting.[76] This will also enable new players to enter the digital market.

Surveillance-based advertisement has far-reaching impacts on people's personal interactions, choices and participation in democratic debates. Measures intended to increase transparency can help to better understand the scale of the issues, but these are not enough to prevent and mitigate the ongoing human rights abuses. The individual and societal harms created by intrusive targeting and personalisation require a systematic response. From privacy intrusions to content curation, invasive tracking harms the right to freedom of opinion in tangible ways. It is a positive obligation of states to protect this absolute right from such interferences by creating an adequate regulatory framework establishing and enforcing strong human rights safeguards.

75 Civil society efforts to gain more transparency have confronted obstacles: https://algorithmwatch.org/en/defend-public-interest-research-on-platforms/.
76 Natasha Lomas, Data from Dutch public broadcaster shows the value of ditching creepy ads, available at: https://techcrunch.com/2020/07/24/data-from-dutch-public-broadcaster-shows-the-value-of-ditching-creepy-ads/?guccounter=1, 2020.

1.2 Guiding note on online targeting

While many intermediaries target their users (as well as non-users) via behavioural profiling and cross-site tracking, online gatekeepers with unprecedented access to large amounts of users' data are leading the adtech industry. In practice, surveillance-based advertising starts with an ad publisher who operates a website or a mobile app that delivers a service or content. They provide a space for placing ads on their platforms and/or access to data about their users. Their trading partners are marketers, companies that are eager to sell their products to the most valuable customers. But third-party vendors and online ad-exchanges stand in between these actors. They operate in the shadows, do not have any direct relationship with users, decide which ads will be placed on which sites, and receive a part of the transaction. This complicated advertising network collects, analyses and merges extensive amounts of personal data without people's knowledge. Neither publishers nor marketers are able to fully or even partly control this process.

Major internet intermediaries assert market dominance across the advertising ecosystem by holding all three roles simultaneously – acting as ad publishers, marketers and third-party vendors. Their dominance is further reinforced by their virtually limitless access to data, including data from their own services and third party data. This creates an enormous power imbalance that fuels unfair competition in the digital market and poses a risk of systemic human rights abuse.

The main focus of this part of the report is internet intermediaries whose business models heavily rely on online targeting. While many intermediaries target their users via behavioural data and cross-site tracking, which amount to human rights intrusive practices per se, these online gatekeepers with unprecedented access to large amounts of users' data are also leaders of the adtech industry. For instance, large social media platforms such as Facebook have developed very granular systems for their advertising interface thanks to which they control major advertising revenue globally. The digital rights organisation Panoptykon mapped and described in depth the adtech ecosystem developed by the gatekeeper Facebook, as well as its impact on human rights. Panoptykon points out that Facebook is not only a passive intermediary between

advertiser and users.[77] It enables advertisers to select criteria that are then interpreted by Facebook's algorithm in order to achieve the advertiser's desired objectives.

This report uses Facebook as a practical example to demonstrate how surveillance-based advertisement works in practice. First, advertisers are able to select their target audience based on targeting criteria that are determined by the intermediary. There are a number of criteria that advertisers can use. Among others,[78] advertisers can choose to target a custom audience or so-called lookalikes. Both criteria were introduced by Facebook in recent years and are often described as follows:

- **Criterion of custom audience** is based on an advertiser's own information that they hold about their users and can upload to the intermediary. Consequently, the intermediary's algorithm matches this information with its own data about the users – without revealing the users' profiles to advertisers.

- **Criterion of lookalike audience** enables advertisers to target a group of users who are similar to the originally desired one. In practice, the intermediary predicts what audience shares the characteristics with the original targeted group – the so called "seed audience". [79] Lookalikes are identified by the intermediary's matching algorithm.

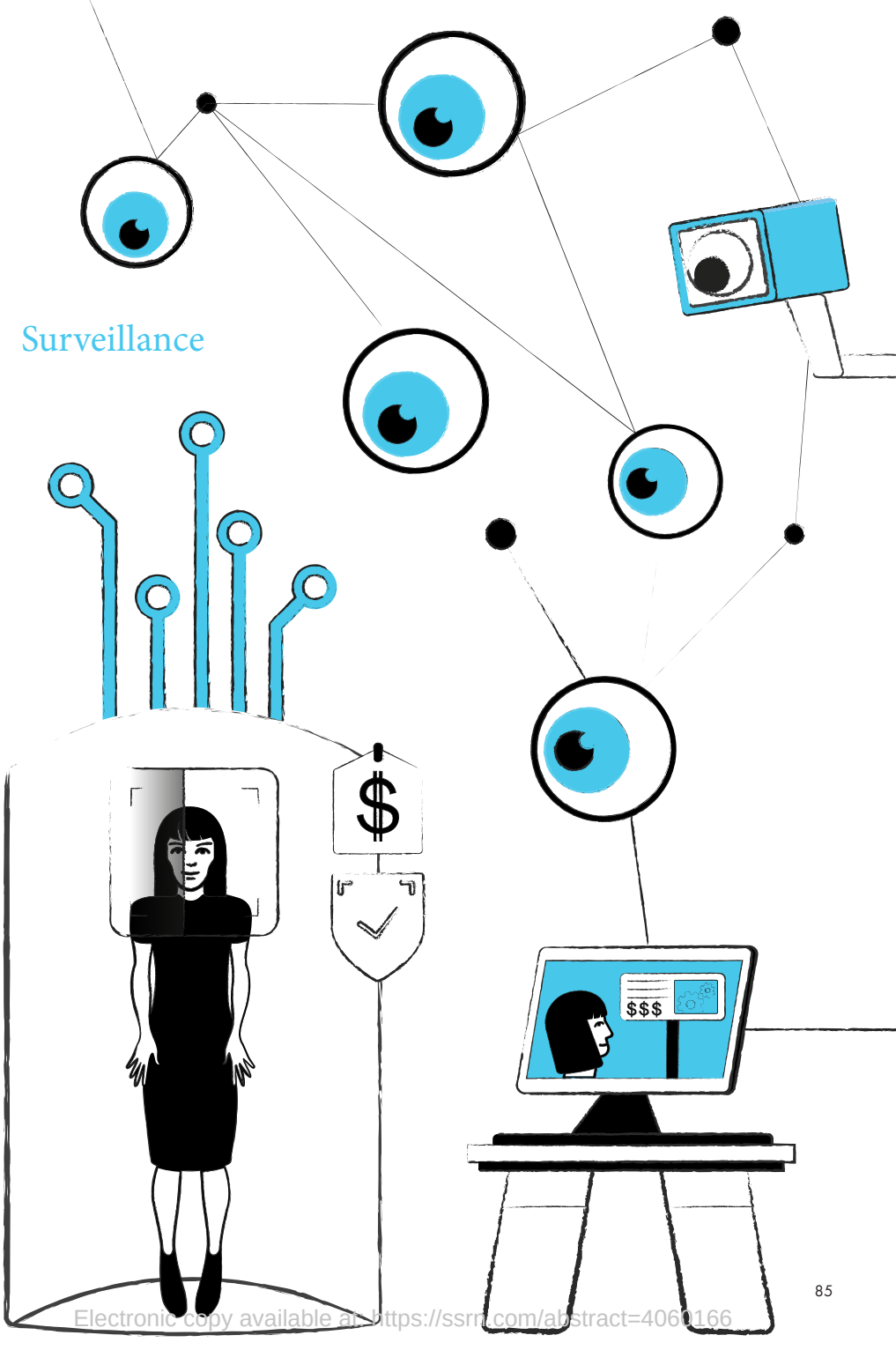
Online platforms are able to target individuals with high precision because they possess data and knowledge about individual users and non-users alike. Big data analysis allows them to predict individuals' behaviour, using data that is directly provided by users or obtained by observing online activity and behavioural patterns of users and others. Highly sensitive algorithms create profiles based on behavioural data— habits, preferences, dislikes and interactions with users. These profiles may even include conclusions drawn from the times that users are most active online.

77 Panoptykon Foundation, Who (really) targets you? Facebook in Polish election campaigns, 2020.
78 Panoptykon Foundation, Who (really) targets you? Facebook in Polish election campaigns, 2020.
79 Panoptykon Foundation, Who (really) targets you? Facebook in Polish election campaigns, 2020. See also, Norwegian Council for Consumer Protection, Out of control: How consumers are exploited by the online advertising industry, 2020.

Both the creation and subsequent use of profiles are privacy invasive and involve assumptions, and they can lead to discrimination. Algorithms are also capable of inferring further information about individuals that the targeted individual does not intend to reveal. The rationale behind this technique is the notion that the more companies know about their users, the greater the likelihood that they can successfully predict and potentially manipulate them. This information is then used to deliver specific content and advertising "at the right time and in the right context," to incite users to buy certain products or services, or to watch certain videos.[80]

80 Vladan Joler, The Human Fabric of the Facebook Pyramid, SHARE Lab Foundation, 2017.

Surveillance



2. Human rights-centred recommendations on regulation of surveillance-based advertisement

2.1 Recommendations to strengthen users' empowerment and personal agency in online ecosystem

- **States should take a people- and user-centred approach to strengthening user empowerment, individuals' agency and control over their data.** There is a significant risk created by our inability to know if we have been profiled or identified, how we have been profiled or identified, and by what algorithm. Yet, transparency alone is not enough for people to control the use of their information. Transparency should be coupled with robust, actionable rights to reject such practices.

- **States should ensure that strengthening users' empowerment and personal agency is not mutually exclusive with constructing complementary systems of external oversight and inquiry.** It is important to prioritise user agency and control in the current socio-political environment.

- **States should invest in research** to develop an empirical foundation for identifying and understanding the effects of surveillance-based advertising on user autonomy and agency. Without further empirical studies, there may be an oversimplification of users' experiences based on scarce data points and research that is focused on particular online communities.

- **States should promote a regulatory framework to improve information distributed to users,** to allow users to exercise free choice in the advertisements they view and to which they respond. The framework should also ensure that users are more aware of the data collected about them and how it is used (including the reasons why a specific user is targeted for a specific advertisement).

- **States should clarify where existing media and content regulation applies to virtual content.** Where gaps are identified, states should review and develop policies and recommendations to moderate online content in the context of the inaccessibility (or "black boxing") of online ecosystems and platforms.

- **States should promote business practices that provide alternatives to current surveillance-based advertisement.** Current business practices of internet intermediaries create a problematic concentration of power that negatively impacts users' personal autonomy and agency.

- **States should ensure private actors act in accordance with the UN Guiding Principles on Business and Human Rights** so that corporate values and governance structures do not prioritise profit maximisation at the expense of human rights and democratic values. The public is increasingly demanding that businesses do not operate in a commercial vacuum but reflect democratic values and priorities.

- **States should encourage the private sector to pursue non-legal avenues to promote greater transparency and accountability.** Private initiatives on codes of ethics play a crucial role in corporate social responsibility. In isolation, however, such self-regulatory approaches alone cannot provide effective protections against the potential for surveillance-based advertising to infringe on the absolute right to freedom of opinion.

Recommendations on outreach and raising awareness about surveillance-based advertisements among the general public

- **States should promote awareness and digital literacy** so that individuals know how to manage their own media consumption and use of internet intermediaries. Users should have a meaningful understanding of why they are shown targeted content as well as how their personal data is being processed and accessed. It is important

for users to understand not only the amount of their personal data that is being processed, but also the type of information that can be accessed, who can access it, and the way certain information can be linked to protected characteristics. Increased digital literacy is important to empower users and strengthen resilience to an ever-adapting industry.

- **States should recognise how surveillance-based advertisement can impact rights to equality and non-discrimination in combination with the right to freedom of opinion and expression.** Surveillance-based advertising creates different and potentially discriminating experiences, both within and between groups of people sharing certain characteristics. This can be exacerbated because certain groups lack digital literacy and that can compound negative experiences with surveillance-based models.

- **States should encourage private actors to consider the concept of "social licence",** which seeks to ensure that private and public service providers act responsibly and ethically in the best interests of the community.

- **States should invest in research** to develop a strong empirical foundation that can ensure outreach and awareness initiatives addressing practical issues of how the public responds to online manipulation and targeted advertising strategies.

Recommendations for legally mandated meaningful transparency: different layers of transparency

- **States should ensure meaningful transparency of surveillance-based advertising.** Personalisation of surveillance-based advertising means that different individuals see different ads based on a number of factors, including time of day, context, demographics, personal characteristics and behavioural patterns. Yet algorithmic systems that are being fed with users' data are profoundly opaque, often described as being "black-boxed". Hence, the decisions behind surveillance-based advertising are near to impossible for users (or regulators) to understand. As a consequence, users lack any meaningful comprehension of why they are being shown a particular

ad at a particular point in time, and how their personal data is shared and used in the process.

- **Designated oversight bodies with expertise in the areas of equality and non-discrimination should be empowered to monitor and address the unequal or discriminatory effects that surveillance-based advertising has on marginalised groups.** States should consider various approaches to accountability for harmful surveillance-based advertising. States should consider the merits of self- and co-regulation models, corporate accountability and governance, litigation mechanisms or alternative e-courts in creating responsibility for adverse consequences.

- **States should ensure that equality bodies are empowered to undertake strategic litigation** to challenge discriminatory outcomes of automated measures.

- **States should collaborate with academia, civil society and independent stakeholders** to refocus transparency efforts on achieving greater access to large-scale disaggregated data that can enable research and understanding of data-driven profiling and advertising. Transparency is required for states and the public to know how surveillance-based advertising is deployed. This will enable meaningful research and allow for challenges to problematic processes. Data access for meaningful public interest research should be based on a legal framework.

Recommendations tackling the interplay between individual versus group privacy

- While international human rights law defines individual rights, online profiling has collective aspects and impacts. Digital profiles are based on inferences and assumptions about a complex web of data and networks. Algorithmic profiling can correlate characteristics and connections to profile individuals from marginalised groups. **States should therefore ensure internet intermediaries respect the right to freedom of opinion and are aware of its important intersection with the rights to freedom of association and expression.**

- **States should consider the constraints of existing legal mechanisms for enforcing collective rights in the context of surveillance-based advertisement.** Current legal systems reflect individual rights, and the only extension to group concerns exists when an individual belongs to a certain group. In these cases, even when an individual makes an informed choice to opt-out of sharing their personal data, they may still be profiled as a part of a wider group that is targeted or categorised by AI systems. States should ensure protection and regulation of the use of personal data, including metadata or demographically identifiable data, which is considered extremely relevant and valuable when it comes to advertising methods.

2.2 Recommendations to develop regulatory and co-regulatory solutions that can effectively address negative impact on human rights stemming from surveillance-based advertising

Recommendations to safeguard the absolute right to freedom of opinion

- **In line with international human rights standards, states should respect and promote the absolute right to freedom of opinion**—which includes the rights to keep one's thoughts and opinions private, to not have one's thoughts and opinions manipulated, and to not be penalised for one's thoughts and opinions.

- **States should emphasise that everyone enjoys the right to hold opinions without interference; the right to seek, receive and impart information and ideas through any media, regardless of physical frontiers; and the right to not be subjected to unlawful or arbitrary interference with their privacy.** States should ensure that individuals can form their opinion while being protected from manipulation by opaque profiling methods that determine when a user is most susceptible to behavioural influence in order to exploit the user's vulnerabilities. Undue influence can stem from practices such as: non-transparent or non-verifiable targeted advertising at scale; tracking and behaviour observation techniques or obfuscating design features

("dark patterns"); or the use of power imbalances to influence thoughts (speed, scale, inaccessibility, "black-boxiness" and systematic non-transparent influence). Such tracking and targeting techniques can lead to self-censorship and conforming effects, which might be more prevalent among certain segments of the population. States should set clear policies and criteria for the line between legitimate influence and illegitimate manipulation based on algorithmic technologies, for which states should consider moratoriums or bans.

- **States should consider the legal recourse for users** to address penalisation imposed by internet intermediaries through flaws in the system, or through user classification, which impacts an individual's online experience regardless of the accuracy of the classification.

- **States should legally ensure anonymity and encryption,** and this should include ensuring that opinions are not disclosed involuntarily.

- **States should invest in digital and media literacy and education campaigns to inform the public about how targeted advertising based on profiling and surveillance methods impacts individuals' online experience, and how surveillance advertising threatens freedom of opinion.** Constant surveillance is not in accordance with human rights, and it risks creating chilling effects on freedom of opinion and expression. States should ensure that individuals have sufficient tools and information diversity to form their opinions freely and to enjoy the positive aspects of freedom of thought and opinion.

- **States should address surveillance-based advertising in the sociotechnical context of content moderation and content curation,** and should identify ways to address the centralisation of power, including due to the bundling of several services.

Recommendations for meaningful transparency and for regulatory measures of online targeting

- **States should develop a human rights policy with emphasis on salient human rights issues**—such as freedom of expression, freedom of the media, privacy and freedom from discrimination.

States are the duty-bearers under international human rights law and hold a positive obligation to protect human rights from interference by others, including by private actors or individuals. States should therefore commit to adhering to international human rights law, and should ensure that national laws and policies regulating internet intermediaries and the advertising industry are fully compliant with the international human rights framework.

- **States should ensure private actors act in accordance with due process and the standards of legality, legitimacy and acceptance of oversight by an independent and impartial judicial body, in line with the UN Guiding Principles on Business and Human Rights.** States should establish a regulatory framework for companies to demonstrate that they have rigorously implemented their responsibilities under the Guiding Principles.

- **States should effectively enforce existing data protection and privacy laws.** In this context, states should provide for principles such as data minimisation and purpose limitation. States should also effectively enforce competition and antitrust laws, as well as other regulations aimed at strengthening human autonomy.

- **States should condition corporate surveillance—including targeted advertising that uses tracking and profiling—on human rights due diligence** and a track record of compliance with the UN Guiding Principles on Business and Human Rights.

- **States should oblige internet intermediaries to provide documentation about AI-based tracking and profiling methods that they deploy for advertising purposes.** States should require internet intermediaries to provide explanations regarding the models used, what data is collected, and for what purpose—as well as performance metrics and testing results for the models used. States should mandate that internet intermediaries must properly explain how their advertising and business models work, how algorithmic decision-making is involved, and how such automated systems make decisions affecting the user. Any disclosure should be made in a way that is understandable and accessible to users. Information on the collection of personal protective characteristics, or proxies of these,

should be included. Additional information should be shared, in a privacy-friendly manner, with researchers and regulators.

- **For any data-harvesting and advertising-based business models, states should require ex ante human rights impact assessments** that are part of a clear regulatory framework, and are transparent, independent and inclusive (involving meaningful consultation with potentially affected groups and other stakeholders). The process should include oversight by a regulatory agency or independent stakeholders with relevant expertise, to ensure the mitigation of adverse impacts of advertising models on prevention of discrimination and preservation of freedom of opinion and expression.

- **States should adopt new constraints, or enforce existing ones, that limit what types of data can be collected, and how it can be used, and what types of data may be disclosed to advertisers, data brokers or third parties.**

- **States should clearly define the way advertising methods cause "harm"** (individually as well as collectively/to democratic processes), based on the precautionary principle, so they can identify a threshold for banning harmful surveillance-based advertising practices. Such bans should include, for instance, forbidding weaponising of sophisticated techniques for influence based on psychological models that assume psychological vulnerabilities and manipulability. For data harvesting for targeted advertising that is within the threshold, states should ensure strict transparency methods – for instance regarding product placement – and human rights due diligence, which puts the best interest of the individual at centre.

- **States should ban indiscriminate mass collection and analysis of user data for targeted advertising that harms users individually or collectively, or interferes with their right to freedom of thought and opinion.** This includes, for example, targeted advertising based on pervasive tracking of users' vulnerabilities or categories of protected characteristics, such as ethnicity, gender, religious belief or sexual orientation. Bans and restrictions of surveillance-based advertising could follow the model of bans on deceptive and subliminal advertising or restrictions on advertising alcohol, tobacco, gambling

or environmentally hazardous materials. Special protection should be considered for vulnerable/susceptible groups, such as children and young people.

- **States should ensure that personalised advertising using scraping of personal data operates on the basis of informed consent and on an opt-in basis.** States should ensure users are able to make choices about which data is collected for which purpose and how they want to engage in online debate and be targeted with advertising (including seeing personalised advertising and being surveilled for advertising in the first place). For less invasive advertising models, at least an option to opt-out of data collection should be provided, and there should be an alternative means for ensuring users' safety online. Consent needs to be explicit, non-coercive and based on informed choice, complying with data protection laws, while acknowledging that advertising models can impact human rights not only by collecting and analysing personal data, but also by using other information and metadata. Users should have control over which data is collected, retained or inferred, and how it is used for advertising. States should promote privacy by design and by default.

- **States should mandate that internet intermediaries must provide information about their revenue model and ensure a network for transparency.**

- **States should oblige internet intermediaries to notify users when they are subjected to any form of tracking and profiling,** to tell users how such mechanisms operate, and to provide for opt-in or opt-out options in an easy and user-friendly way. States should mandate that intermediaries must disclose whether advertising content is shown based on users' own history, location information, social media activities, demographic characteristics or other information (including proxies and "lookalike audiences" that group users with certain characteristics). Intermediaries should also be required to disclose their targeting parameters and audience categories (based on behaviour as well as content), as well as the guidelines against which audience categories are evaluated, and whether algorithmically generated categories are reviewed by human reviewers before being used.

- **States should ensure that users have access to profiling data that internet intermediaries hold about them,** as well as any inferences made about them (including metadata, such as assigned categories and the list of advertisers attempting to influence them). This data should be made available to users upon request, in a comprehensible and accessible format. Users should be able to rectify and delete their profile.

- By introducing certain bans and mandating transparency about what data is collected, stored and analysed, and what kind of advertising decisions the data is used for, **states should address the opaque surveillance-based advertising that may impact the ability of individuals to use internet intermediaries' services as forums for free expression, access to information and engagement in public life.**

- **States should require regular transparency reporting,** mandating minimum requirements on the data collected, categories used and automation involved – and how these impact content and advertising provided. Intermediaries should also be required to provide mandatory, functional advertising libraries.

- **States should put in place a framework for internet intermediaries to disclose their human rights impact assessments and ensure external independent review.** These should include assessments of how freedom of expression and information risks are associated with their targeted advertising policies and practices, as well as assessments of discrimination risks.

- To ensure independent external audits of advertising models, **states should require internet intermediaries to conduct reporting in line with privacy and data protection and is accessible to all relevant public authorities and independent stakeholders,** including researchers and civil society organisations.

- **States should require that internet intermediaries grant researchers and civil society organisations access to their advertising data,** so they can evaluate advertising practices, and their individual and collective impact, and inform public interest-driven research.

- **States should ensure democratic governance,** and recognise and empower designated oversight bodies that have expertise in the areas of equality and non-discrimination to monitor and address unequal or discriminatory effects of surveillance-based advertising on marginalised groups.

- **States should strengthen the independence of data protection bodies** and provide them with sufficient political support and financial resources and competencies.

- **States should encourage multi-stakeholder coordination,** capacity-building and research into the impact of interface design on user behaviour, as well as issues such as "dark patterns". States should also promote research into marginalisation and the gendered nature of digital surveillance and advertising effects – and research of negative externalities of business models based on collection of personal information at a planetary scale, which allows micro-targeting of individuals tailored to their specific attributes, traits and preferences. In addition, research should investigate: targeted advertising's potential to influence behaviour maliciously; the connection between surveillance-based business models and intermediaries' incentive to prioritise harmful content, enabling discrimination harms enacted through algorithmic decision making; and associated chilling effects.

- **States should refrain from arbitrarily accessing data collected by internet intermediaries.** Data requests should be based on legitimacy, legality and necessity, and proportionality, and should provide for judicial oversight. States should adequately enforce safeguards to prohibit mandatory transfer of data, especially to law enforcement, and take specific measures to protect marginalised and vulnerable groups.

- **States should address the concentration of power,** which includes the harvesting of data as a source of market power and further reinforces the dominance of a few dominant intermediaries to the detriment of potential competitors and news publishers. Measures could include, for example, mandating interoperability, data portability (data ownership) through secure mechanisms and/or decentralising power.

- **States should address the impact that the concentration of the digital advertising market has on legacy media and the availability of information of public interest.** States should invest in strong public service media and independent journalism.

- **States should invest in exploring alternative revenue streams that do not rely on commodifying people's private behaviour and do not influence or shape emerging behaviour.** Examples of such alternatives include contextual advertising or targeting according to simple criteria to which a user opts in, direct links between content providers and advertisers without the intermediary ad sector monetising content edited by others (including the media), and efforts to encourage human rights-friendly innovation.

- **States should consider public service platforms** that serve and are fully accountable to the public, based on democratic governance.

2.3 General principles for preventing states from piggybacking on surveillance-based business models

Data-harvesting business models connected to surveillance-based advertising can be abused by states as well. Public authorities increasingly rely on data mining from private companies, which serve as "reservoirs of consumer data". Governments are regularly able to access a variety of data provided by the private sector. In recent years, there were a number of cases reported by civil rights groups documenting how public authorities struck an agreement with data brokers in order to gain access to users' personal data. For instance, the Electronic Frontiers Foundation (EFF) described the case of the U.S. Immigration and Customs Enforcement buying ALPR data[81] from Vigilant to help locate people the agency intends to deport.[82] Such informal agreements between governments and private entities pose a serious threat to the protection of human rights. Unnecessary and disproportionate surveillance may undermine security online and hinder access to information and

81 https://www.techdirt.com/articles/20190321/09165441842/vigilant-customers-are-lying-about-ices-access-to-plate-records.shtml.
82 https://www.aclunc.org/blog/documents-reveal-ice-using-driver-location-data-local-police-deportations.

ideas.[83] Surveillance may create a chilling effect on the online expression of individuals, particularly journalists and members of civil society, who may self-censor for fear of being constantly surveilled. Moreover, surveillance exerts a disproportionate impact on the freedom of expression of marginalised groups, including racial, religious, ethnic, gender and sexual minorities, as well as journalists or human rights defenders.[84] This holds equally true for state surveillance as it does for corporate surveillance.

In particular, advanced development of AI has enabled new possibilities of en masse state surveillance that piggybacks on the architecture of intermediaries' business models. Different forms of content monitoring tools deployed by states to clarify relationships between targeted users, or to assign a meaning or attitude to their social media posts via natural language processing and sentiment analysis, may have serious repercussions for the protection of human rights online. When this process is empowered by machine learning, states can uncover connections and interlinks that are potentially invisible to the human eye. Especially in authoritarian regimes, human rights defenders, political activists and the marginalised may be persecuted for their opinions and views, leading to disproportionate and severe punishments.

This part contains general principles that states should follow, in order to prevent human rights abuse at scale:

1. **Public authorities, and especially law enforcement agencies, should have very limited and specifically targeted access to data, narrowed to specific identifiers or specific categories.**

2. **Data collection by law enforcement should always be based on concrete suspicions.** Law enforcement should only obtain access to specific records and content. No bulk monitoring should be performed, including facial recognition that can enable mass surveillance.

3. **Data collected using special national security powers should not be used for any other government purpose, including law enforcement.** It should be retained for a limited period and deleted once no longer required.

83 A/HRC/23/40, https://undocs.org/en/A/HRC/23/40.
84 A/HRC/29/32, https://undocs.org/en/A/HRC/29/32.

4. **Metadata revealing information, such as who people communicate with, and where and when, can be extremely revealing about individuals' lives, and thus should receive a high level of legal protection.**

5. **Illegal surveillance should be criminalised, with effective remedies.** Illegally gathered data should be inadmissible as evidence, while whistleblowers should be protected when revealing illegal behaviour.

3. Conclusion

This part of the report highlights the impact of AI-enabled and surveillance-based targeted advertising and data-harvesting business models on content curation, information plurality and the ability of individuals to form, hold and express their opinions, and access information freely.

It analyses the link of targeted advertising to the rise of powerful internet intermediaries, who simultaneously act as gatekeepers to expression and information in the digital marketplace of ideas. It also examines how the value and thus prominence of online content is increasingly made dependent on its contribution to generating advertising profit for intermediaries. Illustrating how individuals' data, characteristics and vulnerabilities are exploited for targeted advertising, the report outlines the impact of profit considerations on content governance and online information spaces. The report explores how the current digital ecosystem may interfere with the absolute right to freedom of opinion and with the right to seek, receive and impart information of all kinds, regardless of physical frontiers.

Moreover, the report highlights the link of data-harvesting business models of internet intermediaries to state surveillance. Surveillance creates a chilling effect on online expression of individuals, and journalists and civil society in particular, with a disproportionate impact on marginalised individuals and groups. This holds true for state as well as corporate surveillance.

The report outlines a set of proactive, preventative and responsive recommendations for OSCE participating States. These human rights-

centred recommendations focus on safeguarding the absolute freedom of opinion, on ensuring meaningful transparency, on providing regulatory measures of online targeting, and on general principles for preventing states from piggybacking on surveillance-based business models. While certain challenges, such as the lack of explainability, transparency and accountability of AI-based systems linked to advertising and content governance, need to be addressed urgently, the report also identifies the need to address the broader surveillance-based ecosystem in order to genuinely protect and promote freedom of opinion and expression in the digital age.

OSCE
The Representative on
Freedom of the Media