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

Expert workshop on security
9 April 2021

CONCEPT NOTE

Automated detection tools aimed at potentially illegal content online also referred to by policy makers as proactive measures have been in the centre of academic as well as policy debate. Private actors and policy makers often present artificial intelligence (AI) as a silver bullet solution that in a few years from now, will be able to resolve highly complex issues around the dissemination of potentially illegal content, including the spread of terrorist propaganda. However, these promises that are presented to justify boosting "AI uptake across the economy, both by the private and public sector"[1] disregard the fact that proactive identification, detection and removal of user-generated content carries systemic risks. These risks stem from the automated decision making systems themselves that are deployed by online platforms, and often required by states, either directly through legally binding legislative frameworks or indirectly by increased pressure on platforms to 'do more.'

The goal of this expert group is to first, identify the actual and foreseeable negative impact automated and proactive methods for detecting and evaluating online content imposes on individuals' human rights, with the emphasis on the right to freedom of expression and opinion, as well as on the societal level, including for media freedom; and second, to provide a set of human rights centric recommendations addressed to OSCE participating States with the aim to identify human rights obligations, due diligence standards, and procedural fairness safeguards that can effectively prevent such risks.

The work of the expert group is particularly significant given the large number of legislative proposals to regulate potentially illegal content online that have been recently introduced by legislators across the OSCE region. Therefore, the workshop will convene relevant experts from the field of content governance and regulation of AI to build and identify consensus around rights-respecting regulatory response to the spread and dissemination of illegal content online.

The expert group will not seek to determine what constitutes potentially illegal content. However, given its security focus, it should explore proactive methods for detecting and evaluating:

- **Content that is illegal irrespective of its context:** A typical example of such a content is sexual child abuse material that is prohibited by a number of international legal instruments such as the Council of Europe Budapest Convention, the Lanzarote

[1] European Commission, 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 (COM(2018) 795 final), 2020.

Convention and by Convention 182 of the International Labour Organisation, the UN Convention on the Rights of the Child, and others. However even for this content category, national laws do not provide uniform response.

- **Content that is a part of a wider crime:** For instance, in 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 for the documentation of human rights abuse.

- **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 a content category is 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 will be the offence. It also includes, for example, xenophobia, incitement to discrimination and incitement to hatred.[2]

The expert group should focus on algorithmic commercial content moderation at scale defined as "systems that classify user-generated content based on either matching or prediction, leading to a decision and governance outcome (e.g. removal, geoblocking, account takedown)."[3] In this vein, the expert groups should explore systems that make decisions about content and accounts, including filtering and hash matching that can lead to blocking of repeated uploads of the same content, often with cross-border effect. Notable challenges emerge when AI technologies are deployed to allow for the monitoring by law enforcement of peoples' communications, for instance in social media, under the justification of security and public safety. As a result, individual and group anonymity can be under special pressure. Furthermore, it is not clear what the impact on any illegal conduct or crime of these means may be. This is particularly relevant, as AI technologies are vulnerable to overbroad application of rules they seek to impose and they are context blind, which means that they are prone to generate

[2] This categorisation follows the example set by the work of the Council of Europe's Committee of experts on freedom of expression and digital technologies (MSI-DIG) and its *Draft Guidance Note on best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation.* The Draft Guidance Note will be published in the second half of 2021.

[3] Gorwa R, Reuben Binns R, Katzenbach C, 2020, Algorithmic content moderation: Technical and political challenges in the automation of platform governance.

so-called false positives and false negatives in identifying presumably illegal content online, resulting in arbitrary restrictions of legitimate expressions.

Regardless of technological methods being used, these tools impose prior restraints on the right to freedom of expression and information. In practice, this means that they may a priori exclude persons, groups, ideas, or means of expression from public discourse.[4] There is a strong presumption against prior restrictions of freedom of expression in the international human rights framework as well as in constitutional law, including concerns that 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, child abuse material, and spam 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.[5]

The work of the expert group should provide a clear explanation of the potential negative impact of these tools upon individuals' freedom of expression and the wider societal risks they can potentially carry for freedom of the media, democracy and the rule of law. The group should also address claims around the need to protect public security and safety as justification for the use of proactive identification and detection of potentially illegal content online. Finally, the expert group is tasked with providing technical recommendations that enable identification, analysis and assessment of significant systemic risks stemming from content moderation systems, including when used to prevent the rapid dissemination of illegal content.

[4] Lanza E (2017) National Case Law on Freedom of Expression.Washington, DC: Inter-American Commission on Human Rights.
[5] Llanso E, Hoboken J, Leersen P, et al. (2020) Artificial intelligence, content moderation, and freedom of expression. Trans-Atlantic Working Group Working Papers Series.

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

Expert Workshop on Security
9 April 2021

AGENDA

10:00 - 10:10
Welcome by OSCE RFoM and Access Now

- Teresa Ribeiro, OSCE Representative on Freedom of the Media
    - Welcoming Remarks
- Eliska Pirkova, Europe Policy Analyst, Access Now
    - Introducing the agenda and objectives of the Working group
    - Housekeeping rules

10:10 - 11:00
Tour de table

- Name and affiliation
- What potential negative impact of automated decision-making do you see, concerning measures against illegal content online, upon an individual's human rights, in particular freedom of expression, and the wider societal risk that can potentially follow for freedom of the media, democracy and the rule of law?

11:00 - 11:10
Coffee break

11:10 - 12:00
Session 1: Positive obligations of States to protect individuals' rights against threats against free expression from private actors

What obligations should be put in place to achieve meaningful decisional transparency and accountability in automated decision-making systems (ADM) deployed by platforms in moderating illegal content online?

- Introduction by the Chair covering the main areas for this session
- Discussion among experts

- What transparency obligations should be established by a regulatory framework in order to enable meaningful transparency regarding the specific functionalities of the automated decision making systems that are used in governing potentially illegal content online?

- How can we enable meaningful transparency when it comes to the databases of potentially illegal content that currently remain closed to all stakeholders, including trusted third-party auditors, public authorities and vetted researchers?

12:00 - 13:00
Lunch break

13:00 - 13:50
Session 2: Human rights obligations of States in respect of algorithmic systems that may negatively impact individuals' human rights, and in particular the right freedom of expression

When algorithmic systems have the potential to create an adverse impact on the right to freedom of expression, including the right to seek, receive and impart information, for an individual, for a particular group or for the population at large, including effects on freedom of the media, democratic processes or the rule of law, these impacts engage State obligations with regard to human rights. In general, content classifiers, whether used for recommendation, ranking or blocking may discriminate against content associated with protected categories such as gender or race. How can we secure human rights compliance and achieve algorithmic fairness in ADMs?

- Introduction by the Chair covering the main areas for this session
- How can States ensure that algorithmic design, development and ongoing deployment processes in moderating potentially illegal content online incorporate safety, privacy, data protection and security safeguards by design, with a view to preventing and mitigating the risk of impeding freedom of expression and other adverse effects on individuals and society, in particular for already marginalized voices?
- How can States secure the proper evaluation of datasets used by ADMs in content moderation as well as the functioning of the algorithmic systems that they implement is tested and evaluated with due regard to potential impact on human rights, with emphasis on right to freedom of expression?

13:50-14:00

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

Coffee break

14:00-14:50

Session 3: Public oversight over law enforcement cooperation with private actors in detecting potentially illegal content

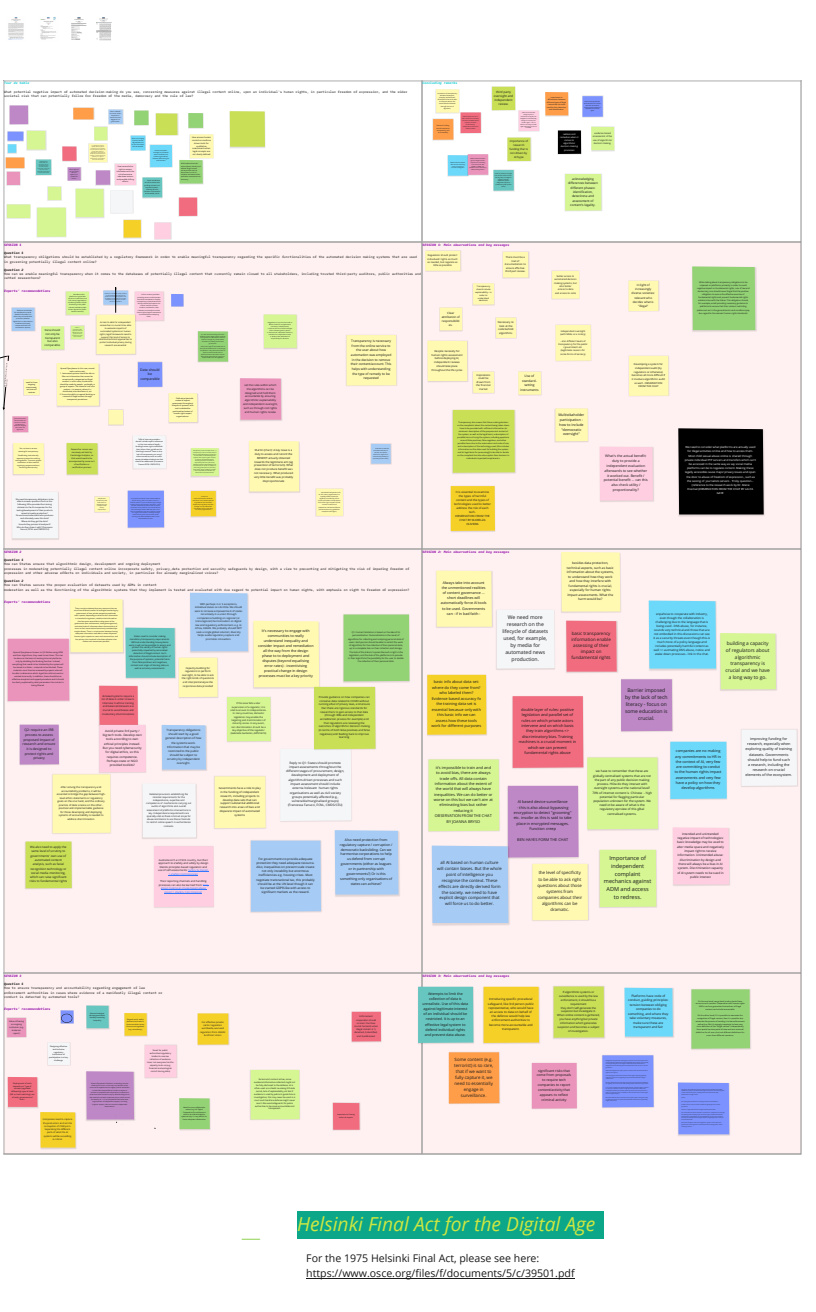
Where rules are unclear, designed in an opaque manner or imposed arbitrarily, neither the scope of the rules nor the consequences of breaking them are known. Furthermore, it is not clear what the impact on the illegal conduct being addressed may be.

- Introduction by the Chair covering the main areas for this session
- How to ensure transparency and accountability regarding engagement of law enforcement authorities in cases where evidence of a manifestly illegal content or conduct is detected by automated tools?

14:50 - 15:00
Coffee break

15:00 - 15:50
Closing remarks:

- Brief discussion on areas not covered by this workshop which would need additional attention (in the scope of the specified subject matter: “security”)
- Summarising the takeaways, seeking to identify technical recommendations from the “how to” discussions in the three sessions
- Explaining the next steps



Helsinki Final Act for the Digital Age

For the 1975 Helsinki Final Act, please see here:
https://www.osce.org/files/f/documents/5/c/39501.pdf