AI *for* Freedom of Expression

"The world must respond to the harm caused by the spread of online hate and lies while robustly upholding human rights" - *António Guterres*, UN Secretary-General

Introduction

In the context of Freedom of Expression, Artificial Intelligence (AI) holds the potential both to protect and undermine this fundamental right.

While AI makes it easier than ever before to create, access, share, and consume content, opening great opportunities for communication and information exchange, it brings complex challenges, like rise of mis/dis/malinformation and hate speech. The factsheet outlines both the benefits and risks of AI for Freedom of Expression and introduces tools for identifying content that is AI generated.

FRAMEWORK

Freedom of expression is protected by Article 19 of both the Universal Declaration of Human Rights and the International Covenant on Civil and Political Rights. It includes three essential components:

1. The right to **seek** information
2. The right to **receive** information
3. The right to **impart** information

Therefore, this right applies not only to producing content but also to accessing and searching for it.

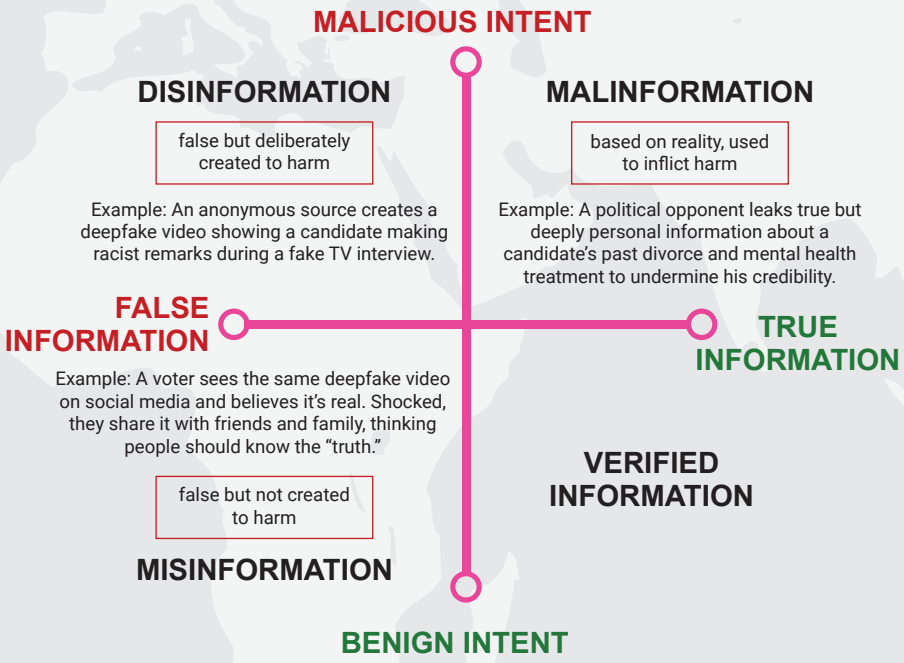
However, freedom of expression is not an absolute right and can be limited, provided those limitations meet three key criteria established by the so-called "three-part test":

1. It must be provided by law
2. It must pursue a legitimate aim
3. It must be necessary and use the least intrusive means to achieve that aim.

That is why some content, usually flagged as "harmful" (e.g. disinformation), can be legitimately removed from circulation.

Challenges

Disinformation, misinformation, malinformation and hate speech existed long before AI came into play. These categories primarily differ in the intent behind and accuracy of the information being shared:



While AI did not create these threats, it has significantly amplified them, making today's information environment increasingly complex, especially during times of crisis or instability when public opinion is more vulnerable to manipulation. This amplification can be explained through **four Vs:**

Velocity — AI enables content, both true and false, to be generated and disseminated at unprecedented speed.

Volume — AI dramatically increases the amount of content produced, which makes it difficult for people to identify truthful information from harmful.

Virality — AI enhances the contagiousness of content, enabling it to spread widely through user interaction.

Verisimilitude — Synthetic content generated by AI appears indistinguishable from real content, making deception more effective.

IMPACT OF AI MIS/ DISINFORMATION ON POLITICAL PROCESSES

2024 witnessed numerous elections around the world and it is during this period, the impact of AI-generated content on political processes became particularly evident. According to research, at least **82 deepfakes targeted public figures in 38 countries between July 2023 and July 2024.** In Slovakia for instance, deepfake audio emerged just before elections, spreading disinformation about electoral fraud, while in Turkiye a candidate withdrew from the presidential race after the release of an alleged deepfake sex tape. This illustrates the gravity and scale of the challenge. While addressing misuse of AI is essential, it must be done with full respect for fundamental rights, particularly the Right to Freedom of Expression.

USE OF AI ACROSS PLATFORMS

While the previous section explored how AI contributes to the spread of information harms and content pollution, it is equally important to acknowledge that **AI tools are now essential in helping platforms manage the enormous volume of content online, from automatically flagging potentially harmful posts to shaping what users see through personalized content curation.** Their efficiency far surpasses that of humans, due to their speed and ability to operate continuously. However, their use also raises concerns, particularly related to bias, accuracy, and transparency.

AI Function	Purpose	Threats and Risks	Instances
Content Moderation	AI agents analyze content, leveraging sentiment analysis, keyword filtering, machine learning and natural language processing (NLP) algorithms to identify harmful/ inappropriate content.	1) AI lacks contextual understanding, sometimes leading to over-censorship (false positives) or failure to remove harmful content (false negatives). 2) AI systems trained on biased data can lead to unfair moderation, resulting in potential suppression of legitimate expression, with marginalized groups being disproportionately silenced.	Social media giants like Facebook, X, and YouTube employ AI agents to detect, flag, and remove content that violates their community guidelines.
Content Curation	AI agents analyze users' behaviors, such as the posts they engage with, the accounts they follow, etc. With this data, AI algorithms can predict the kind of content users like and present it to them in the feed.	1) The algorithmic selection of content is based on intermediaries' policies that follow internal and advertisers' economic interests rather than focusing on accuracy, diversity or public interest (such as news value). This affects the public free flow of information. Engagement is often prioritized over accuracy which fuels disinformation. 2) The use of AI for content monitoring raises questions about user privacy.	Platforms like Instagram and TikTok rely heavily on AI curation to ensure users see content that is most relevant to them. This makes these platforms more engaging and addictive.

SOME AI TOOLS FOR DETECTING SYNTHETIC CONTENT

Fake Catcher (Intel) - Detects video deepfakes in realtime
Microsoft Video Authenticator - Detects video deepfakes
AI Speech Classifier - Marks AI, altered voice content
Optic AI or Not - Determines images are real or AI made
Bot Sentinel - Tracks bot accounts on X
Deepware - Detects deepfake videos
eMonitor+ - Detects harmful content
iVerify - A fact checking tool for stakeholders

UNESCO GUIDELINES FOR DIGITAL PLATFORMS

UNESCO Guidelines for the governance of digital platforms: safeguarding freedom of expression and access to information through a multi-stakeholder approach (2023) recommends that digital platforms comply with five key principles:

- Platforms conduct human rights due diligence.
- Platforms are transparent.
- Platforms make available information accessible.
- Platforms are accountable to relevant stakeholders.
- Platforms should align with international human rights standards in their design, content moderation, and curation.