Cyber protection for human rights activists in key international instruments

Gergana Tzvetkova*

Abstract: There are major threats posed by digital technologies and AI when these are used malevolently against human rights activists. It is worth addressing key recent instruments containing provisions designed to counter such threats.

In 2024, the European Parliament published an in-depth analysis on artificial intelligence (AI) and human rights, which highlighted how AI-powered technology may be used to target, silence, and attack human rights activists/defenders (HRDs):

- Facial recognition technologies 'might be used to identify and suppress political activists or minority groups more aggressively than other populations';
- Governments may improve their capability to identify the content they want to suppress, such as political dissent, social unrest, or information deemed harmful to national security;
- Combining deep packet inspection and AI 'enables the analysis of internet usage patterns, helping to profile individuals' behaviours and interests', which can be used 'to identify and target political dissidents, activists and minority groups';

* Gergana Tzvetkova is a researcher and the co-founder of the Counterintuitive Institute, a Bulgaria-based NGO focused on studying and countering emerging and changing forms of violence and advancing substantive equality, women's rights, and feminist and ethical technologies. With over 12 years of experience in the field of human rights, Gergana has led and contributed to research on gender-based violence, cyber violence against women, gendered disinformation, digital education, and digital rights. Gergana completed the European Regional Master's Programme in Democracy and Human Rights in SEE (ERMA) in 2011.

- AI surveillance and facial recognition tools can be (and have been) used to identify, monitor, and track protesters.

These developments have led to more efforts (by some state and non-state actors and civil society organisations (CSOs)) to adopt and implement various hard and soft-law instruments to better protect human rights activists and limit the negative influence of technology on their lives and work. These initiatives and documents are not perfect – in fact, they have been criticised by some CSOs for falling short on delivering of their promises and the initial expectations. However, it should be recognised that they have established (at least to some extent) a framework for preventing the misuse of technology by illiberal and malevolent actors, emphasising the obligations of technological companies in terms of accountability and transparency, and ensuring the rights not only of human rights activists but also of vulnerable individuals and citizens in general.

There are at least five relevant instruments in this context. In particular, the European Union's (EU) **AI Act** has been heralded as the first-ever legal framework on AI, which addresses AI-associated risks and fosters trustworthy AI in Europe. It pursues a risk-based approach by defining four levels of risk for AI systems: unacceptable, high, limited, and minimal. The following eight practices are associated with unacceptable risk and absolutely prohibited:

1. harmful AI-based manipulation and deception
2. harmful AI-based exploitation of vulnerabilities
3. social scoring
4. individual criminal offence risk assessment or prediction
5. untargeted scraping of the internet or CCTV material to create or expand facial recognition databases
6. emotion recognition in workplaces and education institutions
7. biometric categorisation to deduce certain protected characteristics
8. real-time remote biometric identification for law enforcement purposes in publicly accessible spaces

High-risk AI systems can be on the market only after complying with strict obligations such as being accompanied by adequate risk assessment, mitigation systems, and appropriate human oversight. However, many CSOs – like Access Now, Amnesty International and EDRi – have stated that the AI Act did not go far enough to ban completely the most dangerous uses of AI. Notably, Barkane and Buka support this view:

> The AI Act sets only partial prohibitions on the use of real-time remote biometric identification, predictive policing, emotion recognition and biometric categorisation systems, severely limiting their cope and allowing for a large number of exceptions. Consequently, the use of remote biometric identification and predictive policing systems for law enforcement purposes, although permitted in exceptional cases, could almost always be justified.

The AI Act entered into force in August 2024 and most of its provisions will be applicable in August 2026.

At the Council of Europe (CoE), the first legally binding convention concerning AI – the **Framework Convention on Artificial Intelligence and Human Rights, Democracy and the**

Rule of Law – has been open for signature since September 2024. The document, which 'does not regulate technology and is essentially technology-neutral', has been signed by 15 countries and the EU and is yet to enter into force. The Center for AI and Digital Policy (CAIDP), a prominent non-governmental actor active in the monitoring of AI development and use, has been spearheading a global campaign to support the Convention. CAIDP has worked closely with the CoE on the development of the document and recognised its significance but has urged the organisation to 'make further changes to ensure that the treaty cover AI systems in the private sector and AI systems that may be labelled "national security"'.
The main objective of the 2022 **EU Digital Services Act** (DSA) is to 'create a safer online environment for consumers and companies' in the EU – its rules are designed to:

- 'protect consumers and their fundamental rights more effectively;
- define clear responsibilities for online platforms and social media;
- deal with illegal content and products, hate speech and disinformation;
- achieve greater transparency with better reporting and oversight;
- and encourage innovation, growth and competitiveness in the EU's internal market'.

The obligations by the DSA are especially strict for very large online platforms (VLOPs) and very large search engines (VLOSEs), which means they must identify, analyse, and assess systemic risks that are linked to their services. They must identify and report to the European Commission for oversight risks related to: illegal content; fundamental rights, such as freedom of expression, media freedom and pluralism, discrimination, consumer protection and children's rights; public security and electoral processes; gender-based violence; public health; protection of minors; and mental and physical wellbeing. After that, the companies must adopt measures that mitigate these risks.
The 2024 **EU Directive 2024/1385 on combating violence against women and domestic violence** deserves attention as it contends that 'cyber violence particularly targets and impacts women politicians, journalists and human rights defenders'. Thus, if acts of cyber harassment, cyber stalking, non-consensual sharing of intimate or manipulated material, and cyber incitement to violence or hatred – which is criminalised under the cited Directive – have been committed against a person because that person was a public representative, a journalist, or a human rights defender, then this should be regarded as an aggravating circumstance. EU member states must transpose the Directive into national legislation by 14 June 2027.
The **UN HRC Resolution** '**Human rights defenders and new and emerging technologies: protecting human rights defenders, including women human rights defenders, in the digital age**' was adopted by the UN Human Rights Council on 4 April 2025. The document emphasises

> the particular risks with regard to the safety of human rights defenders in the digital age, including their exposure to unlawful or arbitrary surveillance, unlawful or arbitrary interference with privacy, targeted interception of communications, hacking, including government-sponsored hacking, and all forms of online violence and harassment, intimidation, smear campaigns, threats and doxxing, which disproportionately target women human rights defenders, and measures that prevent or disrupt access to information and communication channels, including Internet shutdowns.

The Resolution calls upon states to adopt measures to protect HRDs and their rights in the digital age. Some of the recommended measures are:

- promoting 'a safe and enabling environment for human rights defenders, including women human rights defenders, to conduct their work both online and offline…';
- encouraging 'diverse and human rights-respecting technological solutions to advance connectivity, including by creating an enabling, inclusive and effective regulatory environment for small, non-profit and community Internet operators';
- expanding 'access to the Internet and secure communication tools, including by increasing funding for such digital security resources as encrypted communication applications and secure reporting channels';
- refraining 'from Internet shutdowns, network restrictions or any other measures aiming to disrupt or prevent human rights defenders from having access to or disseminating information and communicating safely and securely…';
- ensuring that 'biometric identification and recognition technologies, including facial recognition technologies, are not used by public and private actors for mass surveillance, and are used only when consistent with international human rights law and the principles of legality, necessity and proportionality, and also to ensure access to remedies for human rights violations and abuses arising from biometric identification and recognition technologies'.

The cited UN HRC resolution is not legally binding, but it is an important political declaration with strong international support, which introduces concrete measures that states could take to safeguard the work, the rights, and the lives of HRDs.
It is beyond the scope of this blog post to explore the risks that human rights activists face in every country or region or the national legislation and the policies to restrict and target or, alternatively, protect human rights defenders in the digital age. However, the AI and Democratic Values Index published annually by CAIDP since 2021 is an excellent and comprehensive review of AI policies and practices worldwide.