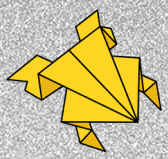


Digital
Security
Lab

HUMAN-RIGHTS-COMPLIANT USE OF ARTIFICIAL INTELLIGENCE SYSTEMS

TOOLKIT FOR CIVIL SOCIETY

Tetiana Avdieieva
Oleksandr Batkhin

2024

This Toolkit outlines the key challenges and risks to human rights faced by civil society, journalists, and human rights defenders all over the globe when applying AI systems. It provides specific examples of threats posed to the security of personal data, equality, freedom of expression, and other related freedoms, as well as elaborates on the ways to avoid or mitigate such risks. The Toolkit is also equipped with a check-list enabling the readers to make a quick verification of their AI system's compliance with basic human rights standards and ensure their choice of technology is informed, careful, and responsible.

Digital Security Lab Ukraine (DSLU) is a non-government organization based in Kyiv, Ukraine. DSLU's mission is to support the implementation of human rights on the Internet by building the capacity of CSOs and independent media to have their digital security concerns addressed and by impacting the government and corporate policies in the field of digital rights.

Contact us:
https://dslua.org
dslua@dslua.org
FB: https://www.facebook.com/dslua
X: @DSLab_Ukraine

TABLE OF CONTENTS:

INTRODUCTION

The growing popularity of artificial intelligence (AI) systems amongst CSOs provides various benefits while also raising multiple concerns for human rights impact and compliance with relevant standards. It is vital to prevent unsafe and unregulated use of AI systems that can lead to various negative consequences for civil society.

Why does civil society use AI instruments? Rapid digitalization equipped all users with various automated tools, starting from outer instruments impacting their rights and ending up with systems consciously used by them to enhance productivity. Civil society is not an exception, willing to gain benefits from AI systems, such as access to fast content creation, effective investigating tools, efficient identification of fake news, and many more. At the same time, they are more vulnerable towards the risks and dangers posed by abusive and irresponsible AI use, which opens the floodgate to a plethora of new risks and challenges. This Toolkit views AI systems in their broadest meaning, introduced by the OECD.

> **AI system** - a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

What are the challenges of irresponsible use of AI? Irresponsible use of AI systems may facilitate the spread of misinformation, data breaches, IP rights violations, algorithmic bias, and other dangerous outcomes. Civil society may subconsciously amplify such risks if their use of AI systems lacks meaningful human oversight, basic knowledge of the technical side, and awareness of the key challenges on the AI market. A diverse set of examples, where unsupervised or poorly managed use of AI has led to human rights violations, includes:

- Data leak revealing that Google-funded AI video generator Runway was trained on stolen YouTube content and pirated films;
- Deep fake video of Moldova's pro-Western president throwing her support behind a political party sharing pro-Russian narratives;
- ChatGPT leaking sensitive user data, supposedly after a hack;
- Samsung employees accidentally leaking trade secrets via ChatGPT;
- Google ex-engineer being arrested for theft of Google's AI secrets for Chinese companies;
- iTutor Group's AI-based recruiting system rejecting applicants due to ageism;
- Healthcare algorithms used by hospitals and insurance companies fail to flag people of color as patients.

What shall CSOs resorting to AI-driven tools commit to? The CSOs that decide to use AI systems should avoid or mitigate all the risks stemming from such tools by adopting the standards of responsible and lawful AI use. Such requirements include, for instance ensuring compliance with international human rights standards and relevant domestic regulations, providing for the safe use of AI systems, their neutrality and transparency, as well as establishing meaningful human oversight, human rights impact assessment, and risk assessment procedures.

The purpose of the Toolkit is to prevent or mitigate the risks stemming from the use of AI systems by civil society actors. Providing the guideline for safe use of both external AI systems (*e.g.* ChatGPT, DALL-E, Midjourney) and internal AI-driven tools (*i.e.* developed, ordered, or adjusted by CSOs), the Toolkit aims to ensure compliance with relevant regulatory standards and human rights requirements, which reflect contemporary standards set by the Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law, as well as the EU Artificial Intelligence Act. The guidelines also provide a checklist enabling the responsible choice of AI systems.

SECTION 1. HUMAN RIGHTS IMPACT ASSESSMENTS AND RISK ASSESSMENTS

A human rights impact assessment (HRIA) is a process for identifying, analyzing, and addressing the adverse effects of AI systems on the human rights enjoyment of any concerned parties (including AI systems' users, other CSOs' members, community members, etc). HRIA and a general risk assessment (RA) are both essential for any CSO that wishes to utilize AI systems' benefits and prevent risks stemming from them. In this respect, it is particularly crucial to avoid substituting HRIA with a mere analysis of risks for organizational models, cybersecurity, financial security, or compliance with domestic laws

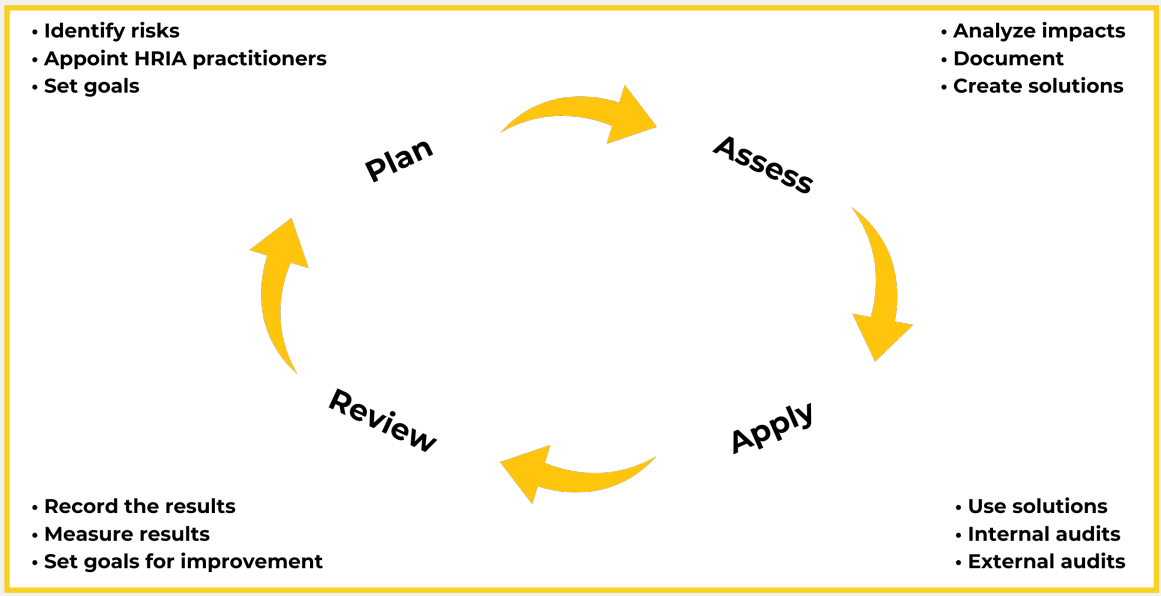
Difference between HRIA and RA:	
HRIA	**RA**
Conducted throughout all stages of AI system life-cycle	
Comprises internal reviews and external audits	
Concerns both positive and negative impacts of the AI system on human rights	Concerns the downsides of AI systems for business strategy, organizational model, cybersecurity, and other aspects of work
Concerns practices around the use of a particular AI system	Concerns the general mode of operation of the organization
Conducted regularly, including after AI system updates, users' complaints, regulatory changes, etc	Conducted regularly throughout all the activities of the CSO

Accordingly, both procedures shall be incorporated into the practice of CSO with adequate methodologies and protocols introduced to ensure the diligent and careful analysis of risks and impacts. It also implies a reconsideration of the system of accountability and reporting inside the organization.

Recommendations. To prevent and mitigate the risks that stem from the AI systems, the CSO shall implement HRIA and RA throughout all stages of the AI system's lifecycle, *i.e.* research, development, deployment, use, end of use, disassembly, and termination. Any CSO that develops or uses AI systems needs to conduct reviews of AI systems' compliance with relevant regulations concerning personal data (PD) protection, freedom of expression, equality, IP rights, etc. In particular, they should take the following steps:

- **Choose the AI systems responsibly.** Before incorporating AI systems into CSO's work, it is best to avoid *prima facie* risks by responsibly selecting such tools. It may include abstaining from the use of certain AI systems, *e.g.* those that manifestly violate human rights. A detailed guideline on responsible choice of AI tools can be found in **Section 7 of this Toolkit (page 23)**;
- **Develop effective methodologies.** For CSO to properly conduct HRIA and RA, there need to be internal methodologies for both. Such methodologies should be up-to-date, accessible to the team and define an algorithm of how the risks and impacts are evaluated, who is responsible for such evaluation, and how the solutions are created;
- **Appoint responsible individuals.** To conduct HRIA and RA, the CSO needs to appoint responsible individuals who will coordinate, conduct oversight over and conduct internal audits and develop strategies of risk mitigation, as well as designated team members who will implement the results of HRIA and RA, mitigating the negative impacts and risks for the CSO;
- **Conduct a thorough and timely RA for the organization.** CSOs should regularly evaluate the risks posed by their operations, for instance financial, legal, security, performance risks, etc, and develop solutions that either remove or mitigate such threats. A continuous RA process should include quarterly reviews of risks and risk management plans as well as yearly reviews of risk management policy, framework, and risk assessment criteria. The intensity of such reviews depends on the levels of risk and there should be additional reviews following the incidents, change of policies within the CSO, or newly identified dangers;
- **Conduct a thorough and timely HRIA for the AI systems.** The CSO should conduct HRIA to review, weigh, and balance the positive and negative impacts of the AI systems and find ways to remove or mitigate the threats. Ensure that internal and external reviews are conducted as early as possible since the start of any new project, *e.g.*, the design or development of an AI system. Regular reviews of the human rights impacts of AI systems should be conducted at least annually with additional reviews in the vicinity of an update or regulatory changes, as well as after any relevant complaint;
- **Conduct internal and external audits.** The CSO should regularly organize internal audits and allow external audits to conduct HRIA and RA by outside independent professionals and by the experts inside the organization, especially where their own expertise in a particular subject does not allow them to meaningfully identify all the risks and impacts;
- **Responsibly choose external auditors.** Depending on the area of conducted RA and HRIA, CSOs should carefully choose the companies for external audits, and review their expertise, reputation, and RA or HRIA methodologies to correspond with the needs, values, and goals of the CSO;

- **Duly implement the outcomes of HRIA and RA.** After conducting HRIA and RA, the CSO should implement all the solutions developed via these processes, removing or mitigating risks and adverse human rights impacts. When implementing the solutions of HRIA and RA, the CSO should always improve its efforts to establish better practices for itself. The full cycle of HRIA and RA then should follow this algorithm:



- **Timely notify developers or suppliers of external AI systems about the identified risks and adverse impacts.** If in the course of HRIA, the CSO discovers risks, adverse impacts, or other issues with AI systems, it should swiftly notify the developer or supplier of such systems regarding those issues so that the developer or supplier improves the situation. If no reaction follows, there are reasonable grounds to consider alternative AI systems;
- **Regularly review the methodologies.** The organization should establish its own review methodology to conduct HRIA properly. For instance, the Danish Institute for Human Rights provides a comprehensive approach to HRIA methodologies, that comprises five main steps:

Planning and scoping	The HRIA practitioners should identify relevant stakeholders to consult throughout the HRIA process. Additionally, preliminary interviews with stakeholders may also take place.
Data collection and baseline development	The HRIA practitioners go into the field to research the level of human rights enjoyment of workers, community members, and other relevant rights-holders. This step emphasizes fieldwork, interviews, and other types of active stakeholder engagement.

Analyzing impacts	The HRIA practitioners should analyze the collected data to identify any relevant human rights impacts and assess their severity. This includes assessing international human rights standards and principles, comparative projects, etc. in line with findings from stakeholder engagement.
Impact mitigation and management	The CSO, HRIA practitioners, and stakeholders should unite their efforts to create a plan for preventing and addressing human rights impacts, prioritizing the most severe of them.
Reporting and evaluation	HRIA practitioners provide a detailed HRIA report that is available and accessible to rights-holders, duty-bearers, and other relevant parties.

- **Conduct compliance exercise.** Ensure human rights compliance within AI systems and their correspondence to applicable domestic and international regulations. For instance, the EU Artificial Intelligence Act establishes requirements of transparency, labeling, human oversight, and risk assessment for those who utilize AI systems. Moreover, the General Data Protection Regulation (GDPR) defines the requirement for explicit consent for the processing of PD. Additionally, it is vital to follow modern unified approaches to HRIA and RA for AI systems, such as HUDERIA, which is aimed to provide clear, concrete, and objective criteria to assess and mitigate impacts on human rights, democracy, and the rule of law.

SECTION 2. PERSONAL DATA PROTECTION AND DIGITAL SECURITY

Data processing plays a crucial role in implementing AI-driven tools into the work of civil society. However, multiple risks emerge both on the legal and technical level, creating potential adverse impacts on human rights, digital security, and the overall effectiveness of the CSO work. Thus, DSLU considers responsible data governance as a necessary precondition for human-rights-compliant use of AI-driven tools.

Personal data protection. The majority of AI systems collect, process, and store PD, creating various threats to user's privacy, which may include sensitive data exposure. According to the AIGS index, around 40% of the countries worldwide also actively rely on AI-based surveillance. Moreover, the growing popularity of facial recognition technologies creates new implications for privacy. As an example

of AI-related privacy violations, in 2023, Clearview AI faced legal action due to the massive scraping of PD from the social media profiles of millions of users without their consent. These technologies create risks of violation of user privacy, covert data collection and storage, malicious use of PD, etc.

Recommendations. To prevent violations of other people's rights, it is important to gather any data regarding them only in a transparent and consensual manner. Any CSO that uses AI systems operating with PD should ensure that the processing of data is consensual, transparent, and lawful. Here are some of the steps that CSOs should take to ensure the protection of people's privacy:

- **Avoid AI-driven tools that resort to manifestly illegal practices.** Avoid AI systems that are known for unlawfully collecting, storing, processing, and training on PD. Additionally, AI systems that do not have clear safeguards in their privacy policies or that resort to data scraping, unauthorized transfer of user data, covert metadata collection, etc, should not be used;
- **Create and maintain PD protection policies.** The CSO should establish clear and transparent policies on PD protection that are available both to the team and to the public. Such PD protection policy should include:
  - Main data protection principles;
  - Data protection strategies deployed by relevant entities including individuals, departments, devices, and IT environments;
  - Rights of the subject of PD;
  - Rules of transfer of PD to the third persons;
  - Pertinent legal or compliance stipulations for data protection;
  - The assigned roles and responsibilities, including data custodians and roles explicitly accountable for data protection activities;
- **Lawfully collect and use PD in AI systems.** Any CSO that uses AI systems that collect PD needs to ensure that those AI systems have a valid legal basis, such as consent, contract, or public interest, to collect and use PD. Moreover, AI systems should provide clear and accessible information to the users about how their PD is collected, used, and shared, and what are their rights regarding that data;
- **Timely and properly delete data.** Using any PD, the CSO should ensure that when the purpose of its processing is achieved, this data is completely deleted and will not resurface through the AI system; parts of the data that may be necessary to retain for reporting or legal purposes should be properly anonymized and securely stored;
- **Lawfully transfer PD to third parties (where necessary).** When transferring PD to third parties, CSOs must comply with relevant GDPR requirements and security policies. For instance, CSOs should enter into data protection agreements with third parties that ensure their compliance with data protection standards, as well as maintain records of data transferring processes;

- **Ensure the rights of data subjects.** It is vital to protect the rights of persons whose data is in any manner processed by the AI system. In particular, such individuals need to be able to:
  - Access their PD, processed by the AI systems;
  - Object to the processing of their PD;
  - Change outdated or inaccurate data;
  - Request the deletion of data;
- **Ensure protection of third persons.** If the CSO collects data via surveillance, face recognition, OSINT technologies, etc (for example, as a part of anti-corruption investigations), it is vital to protect third persons from exposure and unnecessary collection of their PD. For instance, when using pictures acquired by surveillance technologies, all persons other than the target of the investigation should be anonymized by blurring their faces or other identifying features;
- **Avoid using PD in publicly available AI systems.** Since AI systems often store and reprocess data, especially publicly available systems, it is vital to ensure that CSO's employees avoid using PD in prompts or while training an AI system;
- **Regulate access to PD within the CSO team.** To ensure the protection of PD, it is vital to limit access to the PD within the team and allow it only to the designated responsible persons;
- **Adequately and timely notify about data incidents.** In case of any incidents with data, the person who encounters the incident should notify the responsible persons within the team, as well as the concerned third persons or the developers and suppliers of the AI systems (depending on the nature of the incident). Apart from notifying the team about the incident, any employee who encounters data incidents should take appropriate security measures, which are addressed below.

Digital security. Despite its growing popularity, AI may pose major security concerns for NGOs, especially regarding the protection of data. According to the AI cyber security survey of the United Kingdom's Department of Science, Innovation, and Technology, 68% of surveyed companies actively use AI models for their business. At the same time, 81% faced security breaches to their AI or would be unable to identify a vulnerability. The use of AI systems without proper security measures may lead to:

- **Model poisoning:** an attack against an AI model that injects malicious data into the training pool causing generative AI models to produce less accurate results. Sometimes, model poisoning may lead to unprecedented bias, discrimination, or disinformation in the content produced by poisoned AI models;

    - Subset of model poisoning attacks are deliberate: attacks that modify the training data to bias the model towards a specific outcome;
- **Adversarial examples:** modified versions of legitimate inputs that are crafted to fool the model;
- **Evasion attacks:** attacks that bypass security systems by modifying the input data to evade detection or classification by the model;
- **Model stealing:** attacks that involve extracting the parameters or architecture of a trained model to create a copy of the model;
- **Data breaches**: an incident that leads to information being leaked or stolen from the organization. Often stolen data may include sensitive information (*i.e.* PII, commercial and trade secrets, or data pertaining to matters of national security).

The risks that are posed to CSOs by AI models' security vulnerabilities enhance the need to establish adequate security safeguards, implement additional measures, and develop practices to avoid or mitigate security-related risks.

Recommendations. To prevent and mitigate security risks, CSOs shall implement procedural measures on the organizational level, follow the principles of data governance, and clearly distinguish what interactions with AI-driven tools create additional security risks. In particular, CSOs should commit to the following steps:

- **Conduct digital security training (DST) for the CSO teams.** The organization should educate its entire staff on basic principles of digital security, explain and train security protocols, and conduct evaluations to ensure that 100% of the staff is protected from cyber threats and is educated in cybersecurity hygiene. The topics of DST, among other things, can include:
    - Data privacy;
    - Password security and authentication;
    - Shadow IT;
    - Unauthorized software and malware;
    - Email security and anti-phishing;
    - CSO's cybersecurity policies and emergency protocols;
    - Other issues peculiar to CSO work.
- **Check the origin of the AI-driven tools**. Omit AI systems coming from companies or countries with a high index of human rights violations or countries with no sufficient data concerning their human rights compliance, especially those infamous for their surveillance and persecution practices. (e.g. Russia, China, Iran, North Korea, etc.). A detailed guide can be found in **Section 7 of this Toolkit (page 23)**;
- **Develop security policies and protocols**. The CSO shall develop security policies, which, *inter alia*, cover the emergency protocols. The protocols

shall indicate individuals responsible for ensuring data security, as well as mechanisms to protect sensitive data. The emergency protocols should correspond to the CSO's data security policies that would together cover all the risks and responses to them;

- **Personalize the settings of AI-driven tools**. Remove the permit to use PD for further training of the system, deny access to the files on the device, reject the cloud storage of data, and adjust the corporate access depending on the position and responsibility of the individuals inside the CSO team. (*e.g.* ChatGPT allows users to choose which previous inputs can be used to further train the system);
- **Ensure that training and testing datasets are secure and reliable**. CSOs shall necessarily check the reliability of training and testing datasets, ensure that third parties do not have unauthorized access to such datasets and cannot uncontrollably modify their substance to further affect the AI model;
- **Aim to have separate personal and professional devices and profiles**. If possible, maintain separate devices and app profiles for CSO-related and personal activities and try to avoid using them interchangeably for the same purposes. Personal devices or profiles should not contain any data related to the activities of the CSO, especially confidential information. In case maintaining separate devices and profiles is not achievable, make sure to apply strict security measures to personal devices and profiles used for work, including device encryption and other advanced protection mechanisms. Ensure their timely maintenance by an in-house security team or trusted third parties;
- **Ensure protection against phishing and develop incident response mechanisms**. The number 1 rule is to never provide personal or other sensitive data in response to an unsolicited request. Avoid clicking links, files, or attachments if they raise suspicion. Protect your accounts by establishing strong multi-factor authentification. Encourage team members to report any suspicious messages and possible incidents in a timely manner;
- **Adhere to the data minimization principle**. Minimize the use of sensitive data when using AI systems that require it (e.g. AI-based face-recognition technologies, fingerprint biometrics, etc). When using such technologies, CSOs should apply a necessary minimum of consensually acquired and used data that does not create threats if stored or discovered via an AI system;
- **Avoid using sensitive data.** Ensure no use of sensitive data when creating inputs for open AI systems that store the data they're given. Use of sensitive data with publicly available AI models, such as ChatGPT, PerplexityAI, Gemini, and others may lead to this data being either discovered by ordinary users or accessed by adversaries via hacking and data breaches. Therefore, it is better not to use sensitive data with any online AI system that you cannot fully trust to protect it;

- **Introduce anonymization and encryption practices**. Anonymize data by replacing any identifying details with randomly generated ones. Otherwise, when the PD needs to be used, encryption protects it by encrypting identifiable data using secure key material using strong protocols, to be decrypted later. It is important to emphasize that unlike anonymization encryption does not remove the sensitive data as it can be decrypted back, therefore it should be used carefully and the decrypting mechanism, especially the key material which was used should be protected;
- **Regulate the use of virtual assistants.** The use of virtual assistants raises security concerns because as long as the program is installed on the device with a microphone, that program will always pose a risk of background listening and subsequent transmission of sensitive information. Therefore, virtual assistants should be avoided on regular personal or professional devices and used in the vicinity of sensitive information. If this is not possible, limit their permissions to "only while in use" or other the most restrictive settings available on the relevant platform.

SECTION 3. PREVENTING ALGORITHMIC BIAS AND DISCRIMINATION

One of the glaring issues with AI systems is the bias that may occur in the AI systems. Due to either model poisoning, insufficient training data, or simply the unemphatic nature of the AI itself, it cannot automatically filter chauvinism and stereotypes from its outputs. Bias can affect not only the specific uses but the main services provided by the AI models as well. To use the outputs of AI models, there needs to be an additional layer of filtering against bias that they might contain. Moreover, it is important not only to avoid bias in the outputs of the AI models but also to interact with these models in a manner that ensures lower levels of discrimination over time.

Recommendations. It is vital to assess the information that the AI model produces and double-check it to make sure that no bias occurs, as well as ensure that user interaction with the system does not trigger discriminatory outputs. There are several important steps to take to prevent bias on all levels - from personal use of the model to its training and minimizing bias:

- **Avoid manifestly discriminatory tools.** Avoid AI systems developed by companies that are known for discriminatory practices or public positions. Similarly, AI systems associated with scandals and public backlash due to their bias should be avoided (*e.g.* Google's online advertising system displayed high-paying positions to males more often than to women);

- **Filter out biased training data.** When using either publicly open AI systems or ones ordered by the CSO, ensure that training databases do not reflect societal biases and discrimination. Check training data regarding stereotypes and conduct test runs of the AI system to find bias in the outputs. Any detected bias should be eliminated from the database;
- **Follow the data representation principle**. To protect from bias it is vital to ensure that the data is representative and diverse, otherwise, bias will occur by not identifying people of certain groups. If needed, oversampling underrepresented groups or generating synthetic data may be necessary to avoid bias towards minorities, vulnerable and marginalized communities;
- **Adjust AI-driven tools to the local context.** Often AI systems created in certain origins and with certain training data may be unable to adjust their inputs to the local context of the user, especially if the developer does not conduct business in that market. To prevent that the database and/or the inputs need to be sampled with examples that adhere to the local context;
- **Establish anti-bias filters for AI systems before their free access to the Internet**. As the Internet contains unlimited sources of biased training data, it is vital to prevent AI systems from learning via such data by providing them with filters against bias;
- **Ensure human oversight.** To prevent any bias in the outputs the users themselves must adjust the outputs of the AI models to prevent any cultural, racial, gender-based, or any other kind of bias in the text. Regular monitoring of AI systems activities allows the CSO to timely fix biases identified in them. Read more on this topic in **Section 6 of this Toolkit (page 21)**;
- **Draft prompts without biases and discriminatory components.** Due to AI models often being trained on the inputs of the users, the less bias and discrimination will occur there - the better outputs AI model produces. Namely, the more specific description the user provides, the fewer stereotypes the AI model incorporates into the creation of an image or text and vice versa. Therefore, it is crucial to avoid general descriptions and/or prompts that by themselves include bias or stereotypical thinking;
- **Warn about potential biases in the outputs**. When publishing any content influenced by AI systems, disclose the AI-based influence on the content and warn the audience that this influence may cause bias and inaccuracies;
- **Update the system based on complaints regarding bias and discrimination.** If the CSO receives a complaint stating that its AI system is biased or discriminative, it should consider such a complaint, investigate the cause of the incident, and reasonably update the AI system to fix the issue.

SECTION 4. FREEDOM OF EXPRESSION

CSO's dissemination of content generated or modified by AI systems may involve certain threats concerning false or misleading information, confidential information, and transparency with the audience in its interaction with AI systems. For instance, increasingly popular deepfake technology, digital avatars, and chatbots are often used to spread disinformation on a large scale. Unsafe use of these and similar technologies opens the floodgate to disinformation, fraud, leaks of confidential information, etc.

Recommendations. To prevent all of the risks listed above, it is important to ensure the presence of filters, labeling, and verifications for AI-generated content and establish additional mechanisms to process such content. Several steps help to avoid any harm from the AI-generated content, prioritization, and curation of it by AI systems:

- **Develop confidentiality policies.** The CSO should establish and maintain confidentiality policies to provide clear guidelines on handling confidential information. Read more on this topic in **Section 2 of this Toolkit (page 10)**;
- **Labeling of AI-generated/AI-modified content.** Any AI-generated or AI-modified content needs to be labeled so that the audience is never misled. Some AI systems or built-in AI features can automatically provide labeling options, but it is reasonable to always initiate labeling yourself as required by many regulatory acts, domestic legislation, and social media platforms rules;
- **Ensure human oversight.** To prevent any disinformation in the outputs, the users themselves must review the outputs of the AI models to edit factually incorrect or misleading information from the text. Read more on this topic in **Section 6 of this Toolkit (page 21);**
- **Fact-checking.** Any output generated by an AI system needs to be checked on its compliance with the original source and the purpose of its creation. If an AI system quotes any sources, they should also be analyzed before utilizing the system's output in any manner;
- **Disable automated sharing/posting.** Automatic posting on several online platforms (including CSO's website) can lead to the sharing of undesired, unedited, or unfiltered information, which then becomes more difficult to assess on multiple platforms at once. Hence, this feature needs to be deactivated on every platform possible;
- **White-list reliable sources.** Create a list of reliable news and sources that can be used as a comparative indicator of whether any piece of information is factual. For example, several well-known newspapers declaring conformity to the best journalistic standards and practices can be used, such as the

New York Times, The Guardian, BBC, etc. This white list needs to be regularly assessed and updated if any discrepancies occur related to white-listed media. It is also crucial to note that white lists cannot be perceived as grounds to consider information reliable by default. Materials from such sources shall be reviewed, where the slightest doubt regarding the AI input is identified;

- **Media literacy education for the CSO team.** The organization should educate its staff on basic principles of media literacy, explain and train designated staff to assess the credibility of online information, identify and combat disinformation, apply OSINT and other relevant technologies to conduct thorough research, etc;
- **Responsible use of AI systems in investigative journalism.** Any cases of the use of AI systems for investigation purposes should be disclosed to the audience of investigation reports. Moreover, any outputs of investigative AI systems should always be carefully supervised by the human oversight team and double-checked with additional sources that are not AI-based. It is vital to ensure that all the data is publicly available and collected legitimately. Additionally, the investigators should use VPN technology to secure their investigation;
- **Responsible use of content curation tools.** The use of AI tools that provide content curation needs to be adjustable to the needs of the audience and consider human rights. For instance, CSO should provide users with the possibility to modify settings on the websites where the content is shared so that they can prioritize content to their needs or opt out of AI content curation;
- **Responsible use of moderation tools.** When using any content moderation tools, the CSO must ensure that they are effective in tackling all the content (e.g. hate speech, online violence, sexually abusive content, etc) without overly censoring the users' comments and those, who can share their views in the specially designated sections, or exercising bias towards certain views, especially politics and socially significant events. Additionally, the CSO should timely update the AI system itself and its training data so that it can always properly cover new resonant events;
- **Responsible use of chat-bots.** When using chat-bots the CSO must ensure that its staff avoids providing the bot with sensitive information. At the same time, training and developing such AI systems needs to be transparent regarding the used training data and there must be guaranteed filters against harmful information or biases, as well as notification for the users that they interact with an AI and not a real person;
- **Responsible use of digital avatars.** Despite its benefits, such technology opens the floodgate to many issues regarding identity theft and fraud, which is why whenever digital avatars are used, the audience needs to be at all times notified that they encounter an AI-made avatar and not a real person. When

training or developing these AI systems, the use of any personal information either in training or by the avatar itself needs to always be consensual;

- **Deepfake detection.** Apply detection technologies to suspicious pieces of media to verify deepfakes by recognizing vocal or facial inconsistencies, color anomalies, etc. Examples of detection technologies are:
  - **Facial motion analysis.** Detects suspicious patterns in facial animation;
  - **Texture analysis.** Detects suspicious patterns in skin and hair texture;
  - **Audio analysis.** Detects discrepancies in lip sync, audio quality, and frequency;
  - **Metadata analysis.** Verifies the video metadata, e.g. date and time of creation and geographic location;
  - **Error level analysis.** Analyze several frequencies of the video to detect inaccuracies;

  Additionally, any audio information processed by the CSOs should be verified due to the high risk of it being fake as audio deepfakes are the easiest to make. It is reasonable to apply multiple detection tools to cover the broadest scope and manifestations of deepfakes.
- **Use of authentication.** Apply authentication technologies and keep track of inherent indicators in content that can be used to identify deepfakes. Such markers include:
  - **Digital watermarks**: A piece of digital code or image, embedded in the content;
  - **Metadata**: Inherent data that describes a certain file and/or piece of content;
  - **Blockchain**: An open technology that utilizes public transparency as a shield against deepfakes.

SECTION 5. INTELLECTUAL PROPERTY RIGHTS

One of the vital concerns that arise when using any AI-based software is the need to prevent infringement of the intellectual property (IP) rights of both developers of AI-based software and other users. It is especially crucial to ensure compliance with IP regulations if the NGO decides to use unique software in their work. Yet, it is no less vital to ensure the responsible use of publicly available AI systems as it may otherwise lead to infringement of IP rights and create threats of legal repercussions for the organization.

Recommendations. To prevent IP rights infringement while using AI-based software, the most efficient method is to apply software licensing to the models that the CSO wishes to use as it defines the terms of use between the licensee and the software developer. Thus, we recommend to:

- **Avoid AI systems with IP rights violations**. The use of AI systems that by default violate IP rights can open the organization to liability. Therefore, the best solution is to avoid using such systems by responsibly selecting them. Read more on that topic in **Section 7 of this Toolkit (page 27)**;
- **Carefully analyze applicable legislation.** Depending on the country of origin of the AI system, the terms of the license agreement (if applicable), and the home country of the CSO, various domestic laws may apply to the activities of the organization. The main legislation that needs to be taken into consideration includes:
  - International human rights standards;
  - International IP rights acts and standards;
  - Domestic IP legislation for the country of operations of the CSO;
  - Domestic laws of the jurisdiction listed in the licensing agreement or terms of use as the governing one (if applicable);
  - Domestic laws of the country of origin of the AI system developer (if applicable);
- **Consider what is protected by IP.** When assessing any inputs or outputs of the AI system, the types of work protected by IP regulations need to be taken into account. For instance, any original ideas, designs, discoveries, inventions, and creative work produced by an individual or group are protected by IP rights;
- **Adhere to the fair use doctrine.** This doctrine allows the use of copyright content without permission of the rights holder when certain conditions are met. For example, if the use is non-commercial, conducted for research, and results in a new different creation, such an act would most likely be considered fair use. However, it shall be verified in every single case whether the fair use rules apply;
- **Use licensed models.** When ordering or adjusting an existing AI system for use, the CSO should enter into a licensing agreement with the developer to prevent IP rights violations for both parties. The same applies to any outside AI-based software that the CSO uses and that was not developed by the CSO itself;
- **Comply with licensing agreements.** When entering into a licensing agreement, the organization should thoroughly analyze the provisions of such agreement, especially how they determine who holds the rights to the AI system, its outputs, software, data, etc, and who is liable for IP rights violations under the agreement. It is vital to comply with these provisions once the agreement is concluded;
- **Draft inputs in AI models in accordance with IP standards.** The user should appropriately design the prompts to ensure that the input in the model will not lead to the infringing output. Here are two ways to minimize the possibility of that:

Generic approach:	Feeding the AI model as little and as general information as possible.
Specific approach:	Make the prompt original and as specific and detailed as possible.

- **Check compliance of outputs with IP standards.** Two main steps can protect the AI model's output from causing IP rights infringement:

Reference check:	Conduct an internet search for the original piece of content to ensure that no copyright occurred.
Output modification:	Modify the output of the AI model to make the final product a unique piece of content and prevent infringement.

- **Protect the CSO's AI-generated content from third-party claims/ IP infringements.** The best way to protect CSO's content from any legal actions is to use AI systems responsibly and ensure that no IP violations occur while making such content, as indicated above. To protect CSO's content from IP infringements, the organization must publicly indicate that it is the rights holder of that content and the content should be protected by a patent or trade secret, where it is applicable. The same tactics apply to the protection of unique features of the AI system developed fully by the CSO.

SECTION 6. HUMAN OVERSIGHT

Issue: To ensure that the CSO is successful in its compliance with all relevant regulations and standards and that it uses the AI models safely and productively, it is important to ensure human oversight. Any activity that includes the use of AI models needs to include human oversight mechanisms to prevent any harmful effects that AI models may cause.

Recommendations. To prevent any inaccuracies regarding AI models' outputs human supervision over AI models' activity is required. Not only should the users check the outputs before using them, but the organization itself should establish policies, train the team, and constantly monitor the activity of the AI systems, especially if the CSO is training its own AI model or using licensed software. Here are some important steps to take to maintain the needed level of human oversight:

- **Appoint responsible individual(s).** To perform human oversight in an organized manner, it is vital to appoint key members of the team responsible for human oversight and distribute roles of monitoring, evaluation, and decision-making among them;
- **Train the CSO team.** All the team members who are supposed to interact with AI systems should be trained regarding the organization's policies on AI systems and general rules of responsible use of AI. They should be aware of all the crisis protocols and know how to safely use the AI systems, especially those developed, ordered, or adjusted by the CSO. Separate departments within the CSO may be trained to use different systems depending on their needs and sphere of liability;
- **Provide feedback loops.** Establish regular feedback loops within the team regarding the work of the AI model, which would help to adjust the inputs and policies regarding the AI that the organization maintains;
- **Develop crisis protocols.** CSOs should create and periodically update crisis protocols that clearly define the responsibilities of the team during crisis response, and provide instructions on efficient crisis communication and specific tactics for security, technical, reputational, and other kinds of damages. Such protocols should be made always accessible and well-known to the team;
- **Develop a complaints portal.** The CSO that publishes any kind of content should establish a complaint mechanism that allows the audience to report issues and provide feedback to the CSO. The organization needs to appoint staff to monitor and reply to the audience via the complaints portal;
- **Regularly update the system.** The CSO that develops, adjusts, or orders AI systems should conduct regular updates to ensure higher protection and easier, more coordinated human oversight team workflow. Similarly, the use of publicly available AI systems should include regular contact with the developer for updates for the AI system;
- **Regularly conduct human-rights compliance analysis.** Human oversight provides for regular HRIA and analysis of the level of human rights compliance of the organization. The responsible persons should provide regular reports regarding HRIA and successful human oversight. Read more on this topic in **Section 1 of this Toolkit (page 7)**.

SECTION 7. RESPONSIBLE SELECTION OF AI TOOLS

To adequately select the AI tools, the CSO must gather information about them and check if they fit the criteria of safe, low-risk AI tools. For the convenience of the selection process, DSLU developed a comprehensive checklist for each section of this Toolkit that CSOs can use to verify that the AI tool is indeed safe to use. The algorithms of how to use the check-list is the following:

1. For each section answer the questions in the left column with either "**no**" or "**yes**";
2. If your answer fits the conditions of the recommendation (right column), act accordingly to this recommendation.

There are three important notes to take into account when using this checklist:

- The rows for questions and recommendations in each section are colored in either **red**, **yellow**, or **green**, depending on the level of risk. The red row represents an unacceptable risk, the yellow row represents an average risk, and the green row represents a low risk or no risk (just important points to consider);
- To properly assess the safety of certain AI tools it is vital to check and answer questions in all of the sections presented below;
- It is important to note that this checklist is not exhaustive and there may be other issues that require CSO's attention, depending on the nature of the AI tool and the specificity of the CSO's work. The organization must always stay vigilant to emerging risks and use additional compliance instruments together with this Toolkit.

Human Rights Impact Assessments and Risk Assessments.

Question	No	Yes	Recommendation
Is the AI system developed by a company with headquarters in the country with a low human rights protection index (*e.g.* Russia, China, Iran)?	☐	☐	If **"yes"**, this AI tool poses an unacceptable risk and is not recommended for use.
Does the developer or provider of AI systems guarantee compliance with all relevant legal regulations?	☐	☐	If **"no"**, this AI tool poses an unacceptable risk and is not recommended for use until such guarantees are provided (compliance reports disclosed etc).
Whether an AI system is developed with breaches of human rights (such as data scraping or manifestly discriminatory algorithms)?	☐	☐	If **"yes"**, this AI tool poses an unacceptable risk and is not recommended for use.
Does the developer or provider of the AI system have a questionable reputation or is known for mass human rights breaches?	☐	☐	If **"yes"**, it deserves additional attention from the HRIA/RA team and taking steps to ensure human-rights-compliant use of a particular AI system in use.
Do the terms of use contain questionable provisions regarding payments, IP rights on generated content, liability, security, privacy, etc?	☐	☐	If **"yes"**, it deserves additional attention from the HRIA/RA team and communication with the developer to clarify those provisions. If provisions cannot be clarified, it is reasonable to abstain from the use of this AI system.
Do the terms of use provide that the user is held liable for IP rights violations, data breaches, spreading of disinformation, etc?	☐	☐	If **"yes"**, the use of this AI tool places a very strong burden on the CSO and the RA team needs to assess whether it is feasible to use such an AI tool in these circumstances.
Does the developer provide a notification mechanism or complaint portal for their AI tool?	☐	☐	If **"no"**, CSO should establish a mechanism to notify the developer of issues or complaints about the AI system.
Does the developer conduct regular HRIA/RA and/or other processes to avoid or mitigate the adverse impacts of their AI system?	☐	☐	If **"no"**, the CSO's shall exercise due diligence, and conduct more thorough HRIA/RA. If it is impossible, it is reasonable to abstain from the use of this AI system.

Digital security and personal data protection

Question	No	Yes	Recommendation
Is this AI tool known for data scraping, unlawful collecting and processing of personal and sensitive data, or other illegal activities?	☐	☐	If **"yes"**, this AI tool poses an unacceptable risk and is not recommended for use.
Does the AI system developer have a history of addressing security incidents in good faith, and constantly improving their security measures against data breaches?	☐	☐	If **"no"**, this AI tool poses an unacceptable risk and is not recommended for use.
Does the AI system allow for the timely deletion of sensitive information?	☐	☐	If **"no"**, the CSO should make sure that no sensitive information is ever given to the AI tool (for instance, via prompts).
Do the data subjects have the right to object to the use of their PD by the AI tool and, if necessary, access and delete it?	☐	☐	If **"no"**, the CSO must ensure that the use of the AI system is always based on legitimate grounds, being transparent to the data subjects.
Is this AI tool trained on sensitive data?	☐	☐	If **"yes"**, the developer/provider must ensure the CSO that this data was collected lawfully and will not be disclosed within the use of an AI tool. Otherwise, this AI tool poses an unacceptable risk and is not recommended for use.
Does the AI tool allow to personalize settings, *e.g.* prohibit using PD for training, reject cloud storage of data, or reject the access to files on the device?	☐	☐	If **"no"**, this may pose a significant risk unless the developer provides sufficient guarantees of security against data breaches. CSO is well advised to avoid providing any sensitive information to such AI systems.
Can the AI tool gather audio information via background listening anytime by simply being installed on the device?	☐	☐	If **"yes"**, it should be taken into account by the CSO and additional security measures should be in place when using such a tool.
Does the developer/provider of the AI tool have a data protection policy?	☐	☐	If **"no"**, it should be taken into account by the CSO and all relevant issues should be settled with the developer/provider.

Preventing algorithmic bias and discrimination

Question	No	Yes	Recommendation
Is this AI tool known for manifestly discriminatory practices or associated with public backlash or scandals due to its bias?	☐	☐	If **"yes"**, this AI tool poses an unacceptable risk and is not recommended for use.
Was this AI tool trained with unbiased, diverse, and representative data?	☐	☐	If **"no"**, this AI tool poses an unacceptable risk and is not recommended for use.
Does the developer guarantee precautions against bias and or algorithmic hallucinations?	☐	☐	If **"no"**, this should be addressed by CSO through the enhanced moderation of the AI tool's outputs.
Can this AI tool be adjusted to the local context of each country or culture that uses it?	☐	☐	If **"no"**, this AI tool may cause issues for the CSOs that operate within contexts outside of the scope of AI tool's training data. It is better to additionally verify the outputs of this tool or avoid using it.
Does the AI tool have anti-bias filters installed when it has access to the internet?	☐	☐	If **"no"**, there is a high chance that the training data is tainted, and the use of such a tool should be either avoided or sufficiently monitored. Alternatively, the CSO should install such filters itself.
Is the AI tool's training data timely updated, for instance after complaints or incidents?	☐	☐	If **"no"**, new incidents or issues that are not covered by the AI tool will enhance the risk of incorrect outputs, hence it is better to avoid such AI tools.
Can the training data be accessed and adjusted by the users or the CSO?	☐	☐	If **"no"**, the developer shall guarantee that no bias is present in the AI tool and timely update the training data.

Freedom of expression

Question	No	Yes	Recommendation
Does the AI tool provide for automated sharing/posting of its generated outputs?	☐	☐	If **"yes"**, it should provide an option to opt out, otherwise, this AI tool poses an unacceptable risk and is not recommended for use.
Does the AI tool have filters against disinformation and biased or hateful information installed?	☐	☐	If **"no"**, this AI tool poses an unacceptable risk and is not recommended for use.
Does the AI tool provide the opportunity to label AI-generated content as such?	☐	☐	If **"no"**, this AI tool poses an unacceptable risk unless the CSO itself always labels the content as AI-generated or AI-modified, depending on the case.
Does the AI tool establish automated sharing/posting with its work?	☐	☐	If **"yes"**, it should provide an option to opt out of it, otherwise, it causes significant risks and is not recommended for use.
Does the AI tool regularly update relevant data?	☐	☐	If **"no"**, CSO should take additional steps to ensure that the AI tool is up to date. For instance, communicate with the developer/ provider or modify the AI tool (if possible).
Does a content curation tool allow users to adjust its settings according to their preferences?	☐	☐	If **"no"**, such content curation tools should be either updated to allow proper customization or avoided by the CSO.
Does a moderation tool provide an appropriate filter without bias or unnecessary censure?	☐	☐	If **"no"**, the training data or an AI tool itself needs to be updated so that it is effective, otherwise, it should be avoided by the CSO.
Does a chatbot both apply sufficient filters against harmful information and is not trained with sensitive data?	☐	☐	If **"no"**, the information filters need to be fixed or updated, while the developer must guarantee that the use of sensitive training data is lawful. Otherwise, such a chatbot is not recommended for use.
Is the digital avatar trained with PD of the CSO staff?	☐	☐	If **"yes"**, there should be guarantees that the gathering of such data is always consensual and unavailable to the third parties, including the developers/providers of the AI system.
Does the AI tool provide quotations and sources with its generated content?	☐	☐	If **"no"**, the full fact-checking exercise regarding the AI tool's outputs shall be made by a human reviewer.

IP rights

Question	No	Yes	Recommendation
Is this AI tool known for copyright infringement or other IP rights violations of other people?	☐	☐	If **"yes"**, this AI tool poses an unacceptable risk and is not recommended for use.
Is the licensing agreement clear, transparent, and defines all the rights and obligations for both parties?	☐	☐	If **"no"**, entering into such a licensing agreement should be avoided, instead, the CSO may propose alterations to the agreement to fix and adjust the issues.
Does the developer guarantee that no IP rights infringement occurs both during training and within the use of this AI tool?	☐	☐	If **"no"**, there is a high risk for CSO to bear shared liability for IP rights violations, hence such an AI tool is not recommended for use.
Does the developer provide that the rights to the content, generated or modified with this AI tool belong to each individual user?	☐	☐	If **"no"**, CSO must ensure additional modification of AI-generated content that would transform it into original, copyright-protected content.
Does the AI tool timely update after the new IP regulations are introduced?	☐	☐	If **"no"**, the CSO shall analyze the compliance of such an AI tool with the updated IP regulations. If the required level of compliance is not met, it is better to avoid using this AI tool.
Can the AI-generated content be used for commercial, creative, or statutory purposes?	☐	☐	If **"no"**, it needs to be taken into account by the CSO as one of the factors in the selection process depending on the purpose of the AI system's use.

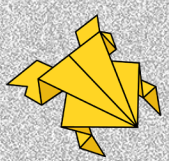
CONCLUSIONS

After considering all the listed recommendations, it is vital to weigh and balance the interests of the CSO when using any AI tools. For instance, in most cases, the ultimate responsibility for adverse impacts caused by irresponsible use of AI systems lies on the CSO itself, even more so if the CSO developed, ordered, or adjusted an AI system rather than uses the publicly available tool. Therefore, it is crucial to apply the relevant standards for AI at **all stages of the AI system's life cycle** - from early development to the use and termination of the system.

Before starting to implement any AI tools into its work, the CSO must carefully evaluate the risks and benefits of the systems, and decide whether potential advantages outweigh risks and dangers. In all instances of applying the AI systems, CSOs shall follow the **basic principles of the responsible AI use**, which are:

- transparency and accountability,
- data security and protection of sensitive information,
- effective and professional human oversight,
- human-rights-centered use of AI-driven tools.

Moreover, the CSO should always **follow the recent updates in the regulatory sphere**, duly implementing the novel legislative initiatives and standards, as well as reviewing the compliance of its practices with the international standards. Finally, one of the key features in the AI sphere is a responsible choice of AI systems, especially those designed for corporate not personal use. When implementing AI tools on the organizational level, CSOs, as actors of vulnerable and risky nature, shall carefully choose the developers and producers of the AI systems and double-check their impact via well-designed HRIA and RA procedures. The progress cannot and shall not be stopped, yet we can make it work for the benefit of civil society by acting diligently and responsibly!



Digital
Security
Lab