



Digital Democracy
Initiative

INTRODUCTORY TOOLKIT FOR RESPONSIBLE AI INTEGRATION



CIVICUS



Acknowledgments

Author

Afef Abrougui (Fair Tech)

Design and layout

Monica Molano

We would like to acknowledge and thank all the civil society actors, researchers, and practitioners who generously shared their time and insights through interviews and expert consultations for this research



Digital Democracy Initiative

The Digital Democracy Initiative (DDI) is a programme to safeguard inclusive democracy and human rights in the digital age. It focuses on supporting local civil society in the Global South, particularly in countries undergoing democratic regression and where civic space is under pressure.

For more information visit:

digitaldemocracyinitiative.net



CIVICUS

CIVICUS is a global alliance of civil society organisations and activists working to strengthen citizen action and civil society throughout the world.

For more information visit:

civikus.org

ABOUT THE TOOLKIT:

This toolkit is based on CIVICUS research on the risks and ethics of AI use by civil society. It is based on insights and input from desk research, the survey responses of nearly 300 CIVICUS network members, interviews and a focus group discussion. It is intended as an introductory guide for civil society actors deploying or seeking to deploy AI. The toolkit supports in navigating major risks and ethical questions associated with AI use, in particular, regarding the right to privacy, risk of bias, and potential for misinformation. This is not a technical toolkit on integrating AI; however, additional resources, technical and non-technical, are listed below.

KEY TERMS AND DEFINITIONS

Artificial Intelligence (AI):

In computer science, AI is defined as the field of computer science concerned with machines and technologies performing tasks previously attributed to humans, such as language processing, making decisions and identifying objects, and analysing data.¹

Training datasets:

Collections of information in different formats—text, images, videos, audio—on which a machine learning model is trained “to make predictions, recognize patterns or generate content”.²

Large Language Models (LLMs):

In the NLP field, LLMs are AI models trained on large text datasets with the capacity to “understand and generate human-like text, recognize patterns in language, and perform a wide variety of language tasks without task-specific training. They represent a significant advancement in the field of natural language processing (NLP)”.³

Data annotation:

The process of labelling raw data such as images, text, videos and audio to make it understandable to a machine learning model.⁴

Fine-tuning:

Databricks defines fine-tuning as “the process of adapting or supplementing pretrained models by training them on smaller, task-specific datasets”.⁵

Human-in-the-loop (HITL):

A process or a system where humans are actively involved in AI workflows so that their feedback can help improve models “to ensure accuracy, safety, accountability or ethical decision-making” of AI systems”.⁶

01

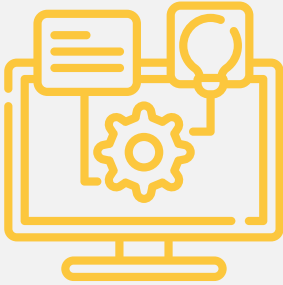


CONSIDER THE PROBLEM YOU ARE ATTEMPTING TO SOLVE AND YOUR ORGANISATION'S CAPACITY

Prior to deploying AI, consider the problem for which the tool will be a solution. Is AI, therefore, an adequate solution to the problem? Can the problem be solved without it? While AI can offer solutions in certain cases, in other cases it may not be an adequate solution, and its use may even backfire by creating new problems or harms.

It is important for organisations to assess not only their own needs and the needs of the communities they seek to serve but whether they have capacity and enough resources to deploy and integrate AI in their work. Are there existing tools on the market that respond to your organisation's specific localised needs? Are existing tools adequately trained in the languages or the contexts in which you work? If you want to develop your own tool, do you have access to high-quality training data that has been ethically collected? In some contexts, lack of electricity, internet connectivity, and computing infrastructure can hinder deployment so consider organisational resources and infrastructure before proceeding with AI.

02



LOOK FOR TOOLS THAT BETTER RESPOND TO YOUR NEEDS AND ALIGN WITH YOUR VALUES

Do your research about existing tools. Make sure the tools you are using align with the mission and values of your organisation. Dominant models owned and funded by big tech companies are based on extraction and exploitation of people's copyrighted work and data, and cheap labour resources in the Global South. Further, these models use significant amounts of energy and water resources, exacerbating the climate crisis. Consider using open-source models. These models have their code, and sometimes also their training data, open to the public which allows others to inspect it, improve it, and use and build on it. This can make it easier, for instance, to identify and fix security bugs or inaccuracies and biases in outputs.

Explore and identify localised models that can meet the language needs of your organisation and which are sensitive to the cultural and social contexts in which you are working. Smaller models (SLMs), as opposed to the generalist Large Language Models (LLMs), are more efficient in specific domains and they are cheaper to develop, use less energy and computing power, and demand less data. Because of their smaller sizes, SLMs can run on local devices and servers, which is better for privacy. For this, though, you need to ensure you have sufficient local storage capacity.



Tip:

Bits of Freedom, a Dutch digital rights organisation, has some recommendations for open source AI chatbots and assistants.⁷

Consider stress-testing different models to decide which one better fits your needs. For instance, you can ask the models controversial or polarising questions or make biased prompts to see how it reacts and what types of outputs it generates. You can also test models to see which ones perform better in the languages you are working or ask questions about the regions and countries in which you are working.

03



DEVELOP GUIDING POLICIES

After you have identified use cases that respond to your needs, you will need to draft AI use policies for your employees, contractors, and partners. Policies should outline:

- ✓• How your organisation uses AI and which use cases are acceptable and unacceptable.
- ✓• Any ethical risks and concerns emanating from the above use cases and measures in place to prevent them and mitigate their impacts when they occur.
- ✓• Privacy protection guidelines for use of personal data in these tools
- ✓• How your organisation will ensure transparency in AI to allow for oversight and accountability, particularly in high-stake decisions that could impact people's fundamental rights. Such measures can include ensuring proper documentation of AI outputs and decisions, maintaining and publishing documentation about their purposes and decision-making processes or informing users about their risks.
- ✓• How people affected by your organisation's deployment of AI can seek remedy in cases where an output or decision affects them. This is particularly important when AI is used to provide support or services to communities and groups of people. They should be easily able to identify which person or department to reach out to and how (email, phone, etc.)
- ✓• Internal procedures for overseeing AI decisions and outputs and who will be accountable for them.



Tip:

Tech nonprofit Fast Forward has an AI Nonprofit Policy Builder⁸, helping organisations generate policies guiding their use of AI depending on their use cases, needs, data they handle and ethical risks. The policy is AI-generated so make sure to review it and edit it to better fit with your organisation's situation and abide by any relevant privacy regulations. The tool can generate policies in different languages.

04



RISK ASSESSMENT AND DATA AUDITS

Depending on how your organisation uses AI and the risks involved, some level of risk assessment may be required. If you are using AI for administrative tasks such as calendar management, and summarising content and taking notes, a risk assessment may not be needed. However, some form of human oversight is almost always required, particularly if you are working in a low-resource language. For instance, you may be summarising a document using a model that does not adequately support Arabic or Swahili which could result in errors or hallucinations.

The scope of risk assessment also depends on whether you are using existing AI models or developing your own. In the earlier case, you do not have control over the model, its training data and the parameters and you cannot audit them. However, you can try stress-testing them to see how they respond to certain prompts and if they generate any inaccurate and biased outputs or hallucinations.

If you are developing your own model, you need to audit the data for quality to identify errors, inaccuracies, and personally identifiable information. To minimise these risks, and depending on the model that you are building, consider using small datasets from credible sources that do not contain personal information or copyrighted data. For example, the Mozilla Data Collective has over 470+ high quality datasets in many languages and which are built ethically.⁹

A number of civil society organisations use data based on content that they or their partners developed to ensure it is of high-quality and free from biases. In cases

where data is scraped from the internet or social media, auditing is of utmost importance as this data contains many biases and may not reflect a diverse demographic. Such data may also need to be reviewed to remove any personally identifiable information. Prior to the data annotation stage, where humans label data for the training of models, it is crucial to ensure that human reviewers do not introduce their own biases in the labelling process. Ensuring different levels of human review and that those working on labelling the data have the right skills and knowledge (for instance, language of dialect knowledge or topic knowledge) is essential.

In addition to auditing the data, you will need to systematically test the model. This should happen prior to its launch in a restricted environment before releasing it for broader use or to the public. These assessments should also happen on a regular basis after releases and every time an issue arises. When developing a model, guardrails need to be put in place to ensure that they do not perform outside of defined boundaries such as not generating racist and biased outputs or make wrong and harmful claims or expose personal information. To test the model, you can start by conducting a contextual analysis of scenarios in which the model can cause harm or perform poorly by generating inaccurate information and clarifying the types of harm you will be simulating. You can then test the model by injecting it with adversarial prompts or attempting to attack its safeguards; observe how it responds and analyse its

performance. Based on this analysis, steps are taken to mitigate risks by fine tuning the model's parameters, limiting the personal information, changing the model's guardrails, and reinforcement learning through human feedback loop.

Human review and feedback loop

Use AI as an assistant, not a replacement of human judgement, decision-making, and connection.

When you are conducting research or looking up information, **never take for granted what an AI chatbot or assistant generates**. Previous studies have shown that **they can generate wrong answers or claims that cannot be referenced to sources**. Always check sources. If you are using AI chatbots to draft materials to be shared with others (such as public reports or grant proposals), do not copy and paste answers, and be aware of potential copyright infringement. Ask the chatbots for the sources, check those sources and reference them. Be aware of AI integration into search engines and other services you are using. Be aware of search results that may be generated by AI. On Google search, for instance, the first search result is often AI generated and the company does not give users any meaningful option to opt out of AI search. An open-source alternative like DuckDuckGo provides more meaningful opt-out options.



Tip: Github has a detailed guide on Red Teaming, “a structured, proactive security practice where expert teams simulate adversarial attacks on AI systems to uncover vulnerabilities and improve their security and resilience”.¹⁰ This technical guide can be useful for your organisation’s in-house team /or external experts as they assess risks.

For high-stake scenarios, robust human review should be ensured. High-stake scenarios are those that may result in harm or a human rights violation or may negatively affect your organisation and its reputation. Examples of high-risk scenarios include deploying AI to recruit employees and contractors or to provide support to excluded groups. This is particularly important when deploying this with marginalised communities such as survivors of gender-based violence, refugees or people living with disabilities. An AI chatbot, for instance, can provide quick access to online resources, but it can never replace the value of human empathy and connection.

The table below illustrates some of the most common use cases for civil society, potential impacts, and how they can be mitigated.

USE CASE	POTENTIAL IMPACT / HARM AND AFFECTED STAKEHOLDERS
Using AI in writing proposals 	→ AI-writing tools can help civil society actors improve the quality of their proposals, summarise calls for proposals and generate submission ideas. However, funders may not appreciate a lack of originality in ideas and writing, resulting in unsuccessful applications. While using AI to proofread a proposal submission or improve its structure is itself not problematic, writing entire proposals or generating entire ideas using Generative AI, can backfire. Check what the funder's policy is on AI use in these cases.
Communication 	→ Use of Generative AI in messaging with the public can help CSOs more rapidly craft messages and content in different formats and publish it. While AI can help with proofreading and language, relying on it as a source of information in communicating with the public can increase the risk of misinformation. This can result in reputational damage and a lack of trust among your audiences. Humans should always review outputs and check them against credible sources.
Analysing data for research purposes 	→ AI data analysis tools and functionalities can help organisations analyse data and identify patterns and trends, but errors can still occur in the analysis, and thus organisations should ensure human review of the outputs and analysis of the data.

USE CASE

POTENTIAL IMPACT / HARM AND AFFECTED STAKEHOLDERS

Providing support services



Some organisations deploy AI to provide support services, for instance, by deploying chatbots that refer people to information and resources, and / or support services. In cases where your target audiences are only interacting with a chatbot to access information and resources, providing an opportunity for them to report inaccuracies, hallucinations, biases or other issues should be enough. However, if you are using AI to provide support services for marginalised communities and individuals, for instance, mental health support for refugees or access to legal support for survivors of gender-based violence, never deploy AI as a replacement for human decision making as this can make people feel isolated and even more vulnerable. While AI chatbots can be used by those seeking support to input initial data, they always offer people the option to go directly to a human. Communicate clearly with people that they are speaking with AI and offer them an opt-out option. Ensure decisions and outputs are properly documented to ensure accountability and that mistakes or biases are properly addressed.

Using AI in recruitment and hiring



While this can make sifting through job applications faster, AI-recruitment tools can replicate existing biases, causing discrimination based on gender, nationality, ethnic background, etc. In addition to the direct harm of discrimination on applicants, it can also result in reputational damage for the organisation and legal consequences. Refrain from using AI if you cannot ensure proper human review of its decisions. If AI is used in this case, it should be used to assist humans in reviewing applications and not making recruitment decisions. Be transparent with job applicants about the role of AI in reviewing their applications and offer proper remedy.

05



PRIVACY CONSIDERATIONS

Privacy is a major consideration in the deployment and development of AI models. First, it is important to note that the dominant commercial models have been accused of infringing on the right to privacy. They scrape massive amounts of data from the internet, including personal data, and without user consent, to train their models. Depending on the type of model and type of subscription a user has, these models use prompts entered by users and conversations with them to further train these models. Read their privacy policies to understand how the data and inputs you enter are being used.

Never input personally identifiable and sensitive information in these models. Consider open-source models that do not train on user data.

Are you developing your own model? Consider the origin of the data sources. Is it scraped from social media or the internet? If so, then you will need to anonymise it by removing any personally identifiable information.

Are you using AI models to interact with audiences or provide services to people? Always ask for informed consent. Make it clear to people what data will be collected, how it will be used, where it will be stored and for how long it will be retained.



Tip:

Regardless of of AI deployment, any organisation handling personal information and sensitive data needs strict data protection policies and measures in place.

Where is the data stored? While AI companies or cloud storage providers can assure you that “your data is secure”, take that with a grain of salt. Even the most privacy-centric companies and organisations can still face cyberattacks. When using certain open-source models, you can have the option of storing the data locally on your device; however, be aware that this can require additional storage capacity and space.

06



CONSULT WITH CONCERNED COMMUNITIES AND EXPERTS

Adopt a multistakeholder approach in the adoption of AI, particularly in cases where human rights could be affected. Prior to adoption, consult with experts to understand potential risks and how to address them. Do not only consult with technical experts on AI and AI ethics but also topic experts who are knowledgeable about the domain in which you seek to deploy AI. For instance, as part of a consultation process to build Nuha, a classifier for gender-based hate speech in Arabic dialects, Jordan Open Source Association (JOSA) consulted with feminist organisations on contextualised terminologies.

When relevant, consult with people who are going to be using the AI model, for example, to access support services. Do they want to use AI? What are their concerns? How do they think or hope it can help them? Could AI deter them from requesting support? Listen to what they have to say and include their input into your organisation's decision on whether to deploy AI or not, and if so, how to do so responsibly.

Depending on the scope of AI deployment and its risks, consider conducting consultations on a regular basis, for example, every time a major update is planned or at different stages of the development process, or when an incident has occurred.

ADDITIONAL RESOURCES

- [Things to know before using AI detection tools](#) – Witness
- [Bye Bye Big Tech Step 5: AI assistants and chatbots](#) – Bits of Freedom
- [Using Generative AI for Human Rights Advocacy](#) – Witness
- [The Playbook on AI for Humanity](#) – Fast Forward
- [Simple, Responsible Ways to Start Using AI in Your Nonprofit](#) – Fast Forward
- [Framework For Meaningful Engagement 2.0](#) – European Center for Not-for-Profit Law
- [Generative AI Guide for Civil Society – Digital Justice Network](#)

REFERENCES

- 1 Notre Dame Learning. “AI Overview and Definitions”. Undated. <https://learning.nd.edu/resource-library/ai-overview-and-definitions/> (accessed 30 April 2026).
- 2 IBM. “What is training data”. Undated. <https://www.ibm.com/think/topics/training-data> (accessed 30 April 2026).
- 3 Hugging Face. “Natural Language Processing and Large Language Models”. Undated. <https://huggingface.co/learn/llm-course/en/chapter1/2> (accessed 30 April 2026).
- 4 Coursera. “What Is Data Annotation in Machine Learning”? 19 December 2025. <https://www.coursera.org/articles/data-annotation-in-machine-learning> (30 April 2026).
- 5 Databricks. “What is Fine Tuning”? Undated. <https://www.databricks.com/blog/what-is-fine-tuning> (30 April 2026).
- 6 IBM. “What is human-in-the-loop”? Undated. <https://www.ibm.com/think/topics/human-in-the-loop> (30 April 2026).
- 7 De Heer, M. & Knaap, M. “Bye Bye Big Tech Step 5: AI assistants and chatbots”. Bits of Freedom. 22 January 2026. <https://www.bitsoffreedom.nl/2026/01/22/bye-bye-big-tech-step-5-ai-assistants-and-chatbots/> (accessed 4 May 2026).
- 8 Fast Forward’s Nonprofit AI Policy Builder. <https://www.ffwd.org/nonprofit-ai-policy-builder>
- 9 Mozilla Data Collective Datasets. <https://datacollective.mozillafoundation.org/>
- 10 Smith, T., Pignati, A., Mitsoulis-Ntompos, P. and Micheal. “AI Red Teaming: The Complete Guide”. Github. <https://github.com/requie/AI-Red-Teaming-Guide?tab=readme-ov-file#what-is-ai-red-teaming> (accessed 4 May 2026).



Digital Democracy Initiative

The Digital Democracy Initiative (DDI) is a programme to safeguard inclusive democracy and human rights in the digital age. It focuses on supporting local civil society in the Global South, particularly in countries undergoing democratic regression and where civic space is under pressure.

For more information visit:
digitaldemocracyinitiative.net



CIVICUS

CIVICUS is a global alliance of civil society organisations and activists working to strengthen citizen action and civil society throughout the world.

For more information visit:
civicus.org