

GIRL EFFECT ETHICAL AI GUIDELINES

**FOR THE DEPLOYMENT OF
SOCIAL AND BEHAVIOUR
CHANGE CHATBOTS**





TABLE OF CONTENTS

1. Foreword	1
2. How To Use This Document	2
3. Useful Definitions	3
4. Overview of Girl Effect’s SBC chatbots	7
5. Gains vs Risks of AI for SBC chatbots	11
6. Best Practice When Developing Responsible AI Chatbots	16
Ten Principles For Deploying Ethical AI	
Phase 1: Promoting Ethical AI During Programme Design	
Phase 2: Promoting Ethical AI During Product Design And Development	
Phase 3: Promoting Ethical AI Post-Launch	
7. Final thoughts	25
Footnotes	27
References	28

FOREWORD

This resource was initially developed for Girl Effect’s internal staff involved in the design, development, and deployment of AI-powered chatbots for girls and young women in countries like South Africa, Kenya, and India, in order to provide them with guidance on the ethical dimensions of this complex field in the face of rapid technological developments.

It has since been adapted and made available publicly, as we hope it will prove useful to other teams grappling with best practice, whether they are new to chatbots/AI, or simply seeking to invest more heavily in ethical AI. The recommendations in this guidance have also been compiled as an implementation checklist, available on request.

This guidance is based on desk research accurate at the time of writing, and builds on the authors’ expertise in digital best practice across data rights and privacy, design, and digital safeguarding, as well as on the insights gathered during convenings with 1300+ global members of the Natural Language Processing Community of Practice. In addition, we also consulted 104 Girl Effect staff members, vendors, and Youth Advisors in Europe, Africa and India, to understand their level of understanding, attitudes, and concerns when it comes to the subject of ethical AI¹.

Because of the nature of Girl Effect’s current engagement with AI, this document is focused entirely on the use of AI, including Generative AI, within community-facing Social and Behaviour Change (SBC) chatbots, specifically those which seek to improve the sexual, reproductive and mental health of girls and young women in Low and Middle Income Countries (LMICs). However, we believe that the majority of recommendations are subject-matter agnostic and will be of use to those working in adjacent fields.

This document is pitched at an ‘intermediate’ level, but its broad audience means we have tried to keep technical language and concepts to a minimum. It does not provide in-depth technical guidance (for example, on bias mitigation - although we have included a substantial appendix with further technical advice), rather it provides a holistic approach to mitigating the many possible risks of AI, with suggestions for team members of all levels of expertise involved across the development lifecycle.

Readers may find that much of the guidance in this document is familiar. Indeed, many of the recommendations that were developed in response to the ethical risks presented by other digital channels (for example mobile web, IVR, apps, or social media) remain not only relevant, but of even greater importance.




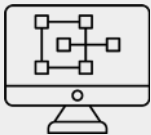


Ultimately, there is no such thing as ‘ethical AI’ - in the same way that there is no such thing as an entirely ‘ethical internet’. Digital interventions all come with an ethical cost, but the potential enormity of the costs, and the number of unknowns, are significant enough that it’s useful to double-down as well as build upon pre-existing efforts to ‘do no harm’.



How to Use this Document

We recommend that both staff and vendors directly or indirectly involved in designing, developing, or evaluating digital SBC tools which use AI in some capacity, read this guidance in its entirety, not least to understand that working towards more ethical AI can only happen if ethical principles are considered and enacted throughout the development lifecycle - and that we each have a part to play in holding each other accountable.

However, we have also broken down the key risks and guidance according to three distinct stages of SBC chatbot development, and used interactive symbols to help readers identify at a glance which section may be most relevant to their role. The table below highlights the activities commonly associated with various roles/workstreams, where you have the chance to make a positive contribution towards rolling out more ethical AI.

	ROLES	RELEVANT ACTIVITIES
	Fundraising And Programme Design	When writing proposals that include AI powered tools, to ensure that considerations relating to required expertise, time and cost of deploying them safely and ethically is built in from project inception.
	Procurement	When writing Reuests for Proposals (RFPs), reviewing Expression of Interests (EOIs)/proposals, assessing the technical infrastructure and data management processes of prospective 3rd party platforms, or interviewing potential 3rd party suppliers.
	Project Management	When planning capacity, activities and budgets, writing reports and updates, summarising activities, or supporting coordination between team members.
	UX & Engineering	When using Gen AI to support in the discovery phase of product design (e.g. desk research and brainstorming; summarising interviews); when designing and building models or tools that use AI; when designing and building data infrastructure or developing processes to manage data.
	Programmes, Content & Safeguarding	When developing Theories of Change for tools which will leverage AI; when developing content for tools which will leverage AI; when using AI to draft content or create images and illustrations.
	Evidence & Impact	When designing MERL frameworks and assessment mechanisms for tools which use AI; when using AI to summarise and/or visualise data which has been collected as part of MERL activities.

By reading this document, we hope that you will come away with:

- A better understanding of key AI terms and a common-ground to build upon.
- A good understanding of the key opportunities and risks that relate to the use of AI in general, with specific examples for SBC chatbots.
- A sense of the risks specifically associated with your own workstream and responsibilities, and useful strategies to mitigate them.
- A sense of the shared responsibility required in order to move towards responsible AI.
- A sense of optimism based on the knowledge that many of the steps required are just an extension of digital best practices.



USEFUL DEFINITIONS

Key terms relating to AI are often debated, even amongst experts. The word ‘AI’ itself has been dubbed a ‘marketing term’², as it’s applied to many different related fields and applications. When talking with colleagues and other stakeholders about the subject, it’s always useful to make sure everyone is on the same page to avoid further confusion, especially when working in new socio-linguistic contexts³.



ARTIFICIAL INTELLIGENCE / AI

Artificial Intelligence is an overarching term that describes the use of computer programs to perform tasks that would typically require human intelligence - from learning, to problem-solving, and language understanding. AI is used as a ‘catch-all’ term, but can encompass any of the following scenarios:

- When you ask ChatGPT to explain the root causes of teen pregnancy in a particular country.
- When Google analyzes your past search habits to recommend results.
- When you use spellcheck, Grammarly or DeepL to assist in writing a report.
- When you use Google maps and follow the traffic-optimized route during a design research trip.
- When a WhatsApp chatbot uses AI to provide a pre-written answer to a user’s question.

NATURAL LANGUAGE PROCESSING / NLP

Natural Language Processing is a subfield of AI focused on language related tasks, and describes the ability of computer programs to understand, interpret and more recently, generate human language in a way that’s meaningful and useful. The types of tasks supported by NLP and relevant to SBC chatbots include using NLP to categorise user questions and signpost them to relevant content, and using NLP to conduct sentiment analysis of self-reported impact data for Monitoring & Evaluation purposes.

PREDICTIVE AI

Increasingly (and not accurately) referred to as ‘old AI’ or ‘traditional AI’, predictive AI is the type of AI most of us were using before the GenAI boom in November 2022⁴. It’s a form of AI that analyzes data (words or numbers), and learns to make predictions, usually in service of a specific task. For example, a predictive model trained on anonymised data can learn to

recognise sentences that indicate a safeguarding disclosure. A system can then be programmed to raise a flag to online moderators or redirect a survivor to appropriate resources.

Although predictive AI is not getting as much coverage because of the excitement over GenAI, one key advantage over GenAI is that the inputs (the data it is trained on) and outputs (the action it performs based on its prediction) are almost entirely controllable by its maker.

Increasingly, AI powered tools use a blend of ‘predictive’ and Generative models depending on the function required, as each model has strengths and weaknesses ⁵.

LARGE LANGUAGE MODELS / LLMs

LLMs are an advanced type of NLP model built using multiple layers and billions of data points and variables. They have the ability to learn and improve from experience, even without being explicitly programmed to do so. LLMs do need human intervention, for example during the preparation of training data, or when correcting errors. Because they learn from such large data sets, LLM outputs often sound very humanlike. ChatGPT, the most widely known generativeAI tool, is powered by an LLM.

LLMs can be adapted, or ‘fine-tuned’, using processes like Retrieval Augmented Generation (RAG) to make them more reliable or more relevant for the intended audience. For example, Girl Effect uses RAG⁶ to make sure UNICEF approved educational content is included by their chosen model as part of the Big Sis sexual health and relationship chatbot.

GENERATIVE AI / GEN AI

GenAI is powered by advances in Natural Language Processing, specifically, Large Language Models. GenAI works by using vast quantities of training data to make predictions about the most likely next word in a sentence, in a split second. When given an instruction or asked a question, GenAI tools can generate entirely new-seeming data, whether text, images or video, that look like they could have been produced by a human.

Generative AI is completely dependent on the information it can scrape from the internet during a particular window in time⁷. It also means that the answers provided by GenAI reflect the biases present in the digital data, reinforcing issues such as language, cultural, and gender bias.

PROMPTS

Prompts are instructions given to an AI powered tool to generate an output, for example, a response to a question. They can be written both by those designing the tool, and those using the tool:

- Prompts are written by those designing AI tools to provide the model with more or less strict ‘guardrails’ in order to improve the quality and safety of its outputs (these are called System Prompts). For example, Anthropic, the organisation behind the GenAI model Claude, makes its system prompt available publicly to provide more transparency on the instructions behind the model⁸.
- Additional prompts are then written by the user of the tool, to elicit the specific information they are looking for. Good prompts should be clear, specific, and involve a degree of iteration (playing around with the prompts a few times to get the best results).

Prompt-engineering is the process of writing and refining prompts. One blind spot in the deployment of GenAI tools for SBC, is that we are not providing users with prompt-engineering skills to increase the efficiency of their use of these tools.

CHATBOTS

The term chatbot has two meanings depending on the context:

- Until recently, chatbots referred almost exclusively to a digital service, usually available via a chat interface on a web browser or instant messaging app like WhatsApp, that enables users to have a human-like conversation via text or voice. Many chatbots are not AI powered, and instead use a pre-determined decision tree architecture that allows users to browse a menu of options (though they may still seem ‘chatty’). These chatbots often incorporate a blend of mechanisms, for example, pre-determined elements, GenAI, and predictive AI⁹.
- Since the advent of GenAI, the term chatbot has been increasingly used to describe GenAI powered virtual assistants such as ChatGPT, Claude or Gemini.

HUMAN IN THE LOOP / HITL

This refers to human oversight or intervention in a mostly automated chatbot service. For example, a user’s question could be handed over to a human agent if the question can’t be answered with confidence by the chatbot. Handing a user back and forth between human and chatbot should be handled sensitively to manage users’ expectations.

In a wider sense, HitL refers to the process of human-led monitoring and re-training that can go into the ongoing maintenance of AI models.





OVERVIEW OF GIRL EFFECT'S SBC CHATBOTS

In this section, we provide some additional information on exactly how and why Girl Effect has historically used chatbots and AI, and how it plans to do so in the future. This information serves as useful grounding for understanding the recommendations in the subsequent sections.

Currently, Girl Effect’s AI-powered tools focus on the thematic area of Sexual Reproductive Health and Rights (SRHR), and mental health, delivering support to adolescent girls and young women via three products:

- Big Sis, a WhatsApp, Moya and Messenger based chatbot for young people in South Africa.
- Bol Behen, a WhatsApp and Messenger based chatbot for girls in India.
- WAZZII, a WhatsApp chatbot for young people in Kenya¹⁰.

These three services allow users to interact conversationally in order to browse information, test their knowledge via quizzes, get answers to questions, or be signposted to useful services, including through human-in-the-loop interaction with a trained healthcare advisor.

The table on the following page explains how each chatbot does or doesn’t use AI, which type of AI, and where humans intervene in the mostly automated process.



CHATBOT	LANGUAGES	USE OF AI	USE OF HITL
Big Sis	South African English	<ul style="list-style-type: none">Core architecture is a pre-defined 'decision-tree'.User questions are classified using a 'predictive' BERT model, providing users with likely content categories that match their question.Keyword matching is used to detect safeguarding disclosures.A proportion of users receive GenAI answers as part of an ongoing study of GenAI vs classification.	<ul style="list-style-type: none">Human agents answer user questions and/or signpost to appropriate services.Safeguarding cases are referred to appropriate services and reviewed by internal Safeguarding teams.
Bol Behen	Hinglish (Hindi+ English)	<ul style="list-style-type: none">Core architecture is a pre-defined 'decision-tree'.Uses LLM models (ChatGPT 4o embedding and generative) for input classification.	<ul style="list-style-type: none">Human agents answer user questions and/or signpost to appropriate services.Safeguarding cases are referred to appropriate services and reviewed by internal Safeguarding teams.
WAZZII	Sheng English (Swahili + English)	<ul style="list-style-type: none">Core architecture is a pre-defined 'decision-tree'.Keyword matching is used to detect safeguarding disclosure.Integration of NLP planned for 2025.	<ul style="list-style-type: none">Human agents answer user questions and/or signpost to appropriate services.Safeguarding cases are referred to appropriate services and reviewed by internal Safeguarding teams.

As laid out in Girl Effect’s AI and ML Vision paper, Girl Effect will be exploring ways to leverage the possibilities offered by GenAI and LLMs, whilst making sure they remain reliable and relevant to often vulnerable end users, using techniques such as:

- Retrieval Augmented Generation
- Prompt Engineering
- Embeddings
- Evaluation framework

This will hopefully result in chatbots that retain their safety and local relevance, but that are infinitely more usable because of the increased nuance and personalisation of their responses in relation to girls’ specific questions, but also in terms of their ability to handle conversational elements such as small-talk.

Girl Effect may soon be using GenAI outside of its community-facing interventions. For example, in the back end, LLMs will be used to suggest categories for user messages which the current models are not succeeding in understanding, at a rate much faster than humans. These rough takes can then be used to underpin further user analysis or fine-tuning of the models.



There may also be increased use of GenAI to support desk-research or administrative tasks, including for proposal writing or HR purposes.

Girl Effect and similar organisations will likely explore using AI as part of its Evidence & Impact (Monitoring & Evaluation) activities. For example, AI could be used to:

- Conduct sentiment analysis, topic modelling or other data analysis to measure the impact of programmes,
- support the translation, transcription and labelling of stakeholder interviews,
- summarise interview or focus group discussions,
- generate draft logframes and indicators.

Another area where AI may become institutionally useful to Girl Effect and similar organisations is in Knowledge Management - International Committee of the Red Cross (ICRC) and the World Bank for example have experimented with using AI to improve their usage and knowledge retrieval of internal documents & data¹¹.





GAINS vs RISKS OF AI FOR SBC CHATBOTS

For readers familiar with the Gartner hype cycle, two years on from the release of ChatGPT, we find ourselves hovering somewhere between peak ‘hype’ and peak ‘panic’ - perhaps approaching the ‘trough of disillusionment’. In this section we examine both ends of the spectrum. The verdict on AI’s utility vs danger sits somewhere in the messy middle.

WHY THE EXCITEMENT?

AI, and especially GenAI, can potentially provide gains across efficiency, quality, consistency, and cost-effectiveness compared to the same tasks carried out by humans, or by non-AI automated chatbots.

These gains are often hypothetical, with very little solid evidence providing a reliable, direct comparison between human vs AI led activities, for example. However, more studies and anecdotal insights are emerging every day, including Girl Effect’s own beta test results comparing GenAI vs predictive AI when answering girls’ questions, which showed that users were 11% more likely to recommend Big Sis to a friend when provided with a GenAI answer, and 11.87% more likely to return for a subsequent visit.

The table below provides example use cases, many of which are already being used or are planned by Girl Effect and similar organisations¹², and for which we expect more and more evidence (either proving or disproving the potential gain) to emerge in the near future.

USE CASE	POTENTIAL GAIN
Girls ask a question to a sexual health chatbot, and a response is instantly provided by a GenAI model.	<ul style="list-style-type: none">• Users who previously might have waited for days, weeks, or forever for an answer, feel supported and heard.• Answers can be provided at a fraction of the cost compared to the number of humans required to answer the same volume of questions.• With the right guardrails, the quality of answers can be more consistent than those provided by humans, who may introduce variations in quality as a result of expertise, literacy, or personal approach.
GenAI is used to handle all the ‘casual’ elements of a chatbot conversation, from onboarding, to small talk, to navigational challenges.	<ul style="list-style-type: none">• The overall user experience of the chatbot is improved, increasing enjoyment and trust by users.• Engagement and conversion rates are improved overall, driving more users, more effectively, to meaningful content.
A predictive model is used in the backend to flag when a user’s question indicates they may be at significant risk of harm, and their question is parsed for additional contextual details.	<ul style="list-style-type: none">• An appropriate, sensitive, automated response can be provided immediately, with a greater likelihood of relevance than via keyword recognition alone.• The user’s question can be prioritised by human experts and signposted more accurately to referral services.• Human experts can use parsed information to avoid a user having to make a disclosure.
Predictive AI could be used to segment users based on their questions and demographic data, and curate an experience unique to their needs and interests.	<ul style="list-style-type: none">• User experiences are improved by providing them with content and services with more expediency.• User retention is increased.• Impact on relevant knowledge, attitude and behaviour is more immediate, and more likely to be reinforced long term.
A LLM is used in the backend to suggest pre-written content to Human in the Loop agents answering user questions.	<ul style="list-style-type: none">• Staff save time by not answering the same questions repeatedly.• Staff can prioritise answering more complex and nuanced questions.

The example use cases in the previous table represent the tip of the iceberg in terms of the possible gains for digital SBC programming in general. For example, GenAI could be used to more rapidly extract insights from behavioral science research, or to generate low-cost visuals that could be used for campaigns. Results from Girl Effect’s GenAI tests within Big Sis also suggest that the cost-benefit of GenAI integration is favorable: whilst costs are higher compared to ‘traditional’ AI-driven approaches, they are not prohibitively so when compared to the improvement in outcomes¹³.

WHAT ARE THE RISKS?

Some of the risks associated with AI use are tangible and immediate - such as the impact of a ‘hallucinated’ answer on vulnerable users. Others are more abstract, and therefore more pernicious - such as the effect of ‘dehumanisation’ as we become more reliant on machines.

The table below lays out the principle ethical risks associated with the use of AI for SBC chatbots, and also indicates whether the risk is primarily associated with GenAI, ‘traditional’ AI models, or both. The risks have been organised alphabetically - assessing the relative importance of each risk is highly subjective ¹⁴.

RISK AREA	EXAMPLE SCENARIOS AND ASSOCIATED HARMS	GEN AI vs TRADITIONAL AI
Bias The data used to train AI models is inherently biased as it reflects the demographic composition of global internet users and AI developers (white, English-speaking, male, Western) - its responses or decisions often therefore reflect biases related to race, gender, age, sexuality, dis/ability, etc. ¹⁵	<ul style="list-style-type: none">A chatbot user asks a question about contraception, and the chatbot answers with an implication that contraception is a girl’s responsibility, reinforcing harmful beliefs.A backend system making recommendations for user content based on their demographic profile may limit their exposure to potentially useful information.	Both - but significantly increased by GenAI.
Data Privacy Tech companies are using our data in order to build and maintain LLMs, enriching themselves and exerting political and market power in the process. The amount of data gathered, how exactly it is used, and by whom, is often opaque, and sometimes unknown even by those who control the LLMs ¹⁶ .	<ul style="list-style-type: none">An AI assistant used by telehealth agents absorbs all user conversations, including personally identifiable information, to further train and improve its model. In principle, employees at the company with the right permissions could access this data, or the company providing the services could use or sell the data for other, unexpected purposes.Chatbot users providing their consent during onboarding will almost definitely not be doing so with a full understanding of what happens to their data¹⁷.	GenAI
Dehumanisation By replacing previously human-led activities or processes with AI, we are losing touch with the intangible but profound benefits that come from human-human interaction, as well as with the more subtle insights that come from applying our own emotional intelligence to problem solving or research ¹⁸ .	<ul style="list-style-type: none">A survivor in need of human contact as part of their reporting or recovery process can feel abandoned or re-traumatised when their request for help is handled by an AI agent.Researchers who have relied excessively on AI tools for analysing qualitative experiences ‘lose-touch’ with the human stories behind this data, leading to a reduction in research quality, and even passion for advocacy.	Both
Digital Divide AI powered tools require more energy and potentially more airtime to run. This could contribute to a widening of the digital divide, which itself contributes to gender inequities.	<ul style="list-style-type: none">Users who previously might have benefited from an on-rails chatbot, find their phone battery or data is drained by using an AI powered one, missing out on useful information and support.	Both


RISK AREA	EXAMPLE SCENARIOS AND ASSOCIATED HARMS	GEN AI vs TRADITIONAL AI
Environmental Harms AI models require significant energy and natural resources to train, run and maintain. This includes the electricity and water required to run (and cool) data centers, and hardware leading to electronic waste ¹⁹ . There is also evidence that the stress and displacement caused by climate events leads to increased incidences of VAW/VAC.	<ul style="list-style-type: none">• The climate crisis is exacerbated, drawing resources away from key SBC issues such as adolescent or maternal health.• Climate related events may reduce access to affected populations, making it harder for development and humanitarian organisations to support them.	GenAI ²⁰
Inclusion & Participation Related to relevance and bias. The models used by GenAI powered tools have, for the most part, been developed without input from those they are interacting with (for example, Gender Based Violence experts, or survivors), leading to the models’ responses or decisions not reflecting specific needs and realities. ²¹	<ul style="list-style-type: none">• At a systemic level, a lack of representativity can lead to the disenfranchisement and marginalisation of already vulnerable groups.• If groups feel that AI tools do not reflect their needs, this will erode trust and turn people away from using tools that might otherwise bring real benefits.	Both
Inequity Emerging evidence suggests that the use of AI as a work tool can either make mediocre outputs appear outstanding, or widen the gap between outstanding and poor performances. There is a fear that using AI is anti-meritocratic, and will enhance disparities already present globally.	<ul style="list-style-type: none">• Those with disproportionate exposure to the digital literacy skills, and access to reliable and affordable electricity, internet, that using GenAI requires, will be more likely to reap the gains of using it.	GenAI
Power & Patriarchy AI is mostly developed and controlled by white men in the global north; as such its success can only perpetuate existing gender and postcolonial power imbalances.	<ul style="list-style-type: none">• By opting to use commercial models which enrich certain demographics, we are participating in the entrenchment of existing unequal power structures.	GenAI
Reliability AI tools are programmed to take a confident tone. This means they may provide answers or analysis that seem convincing, but are actually wrong. These are referred to as “hallucinations.” Whilst developers can work to minimise and detect inaccurate responses, ultimately GenAI is a ‘wild horse’ whose answers can’t 100% be controlled by developers. But - the same is true of humans! Similarly, even where responses are still developed by humans, AI which simply signposts users in a specific direction can also be misleading.	<ul style="list-style-type: none">• A GenAI powered chatbot used to answer mental health questions provides users with dangerous advice²².• A chatbot using a classification model fails to interpret a user’s question, and panics them by referring them for mental health counselling.	Both

RISK AREA	EXAMPLE SCENARIOS AND ASSOCIATED HARMS	GEN AI vs TRADITIONAL AI
Relevance Related to bias. AI tools will provide answers based on their training data, and if this training data does not reflect the reality of the person using it (including language), the response provided or action taken will be, if not incorrect, then less applicable and less appealing to the user.	<ul style="list-style-type: none"> • An AI-powered chatbot that uses voice to text fails to interpret the contextual nuances in regional dialect, reducing the quality of the eventual response. • A chatbot developed to provide sexual health advice provides advice that, whilst not inaccurate, is presented in a way that fails to take into account girls’ lived reality and belief systems. 	Both
Toxicity Similar to reliability, but in this case the concern is that answers provided by an AI model are not wrong, but their tone or content is unwittingly harmful to the user.	<ul style="list-style-type: none"> • A GenAI powered segment of a chatbot, responsible for handling ‘small talk’, responds in an excessively upbeat or insensitive manner to a user expressing unhappiness. 	GenAI
Workers’ Rights AI models require human labour as part of the data preparation and evaluation phases. The work is often menial, repetitive, poorly paid, isolating, and unreliable. Tasks are often conducted in isolation from the wider development process, with limited opportunities for education or upskilling.	<ul style="list-style-type: none"> • By using AI-tools which aren’t transparent about the pay, conditions, contracts, management, and representation of workers in their value chain, we may be perpetuating abusive, dehumanising working conditions and global economic inequities²³. • Junior staff members or other individuals hired to prepare / label data which includes sensitive disclosures can themselves become triggered or burnt out emotionally. 	GenAI

Many of these issues are not unique to AI, nor even to the use of digital tools, and rather reflect historical, complex inequities related to global economics, gender imbalance, racism, and the business of development work itself. It can be useful to reframe the risks of AI as ‘a new manifestation of old problems’ - and as a result remind ourselves to revisit existing principles and approaches to minimising harm.



BEST PRACTICE WHEN DEVELOPING RESPONSIBLE AI CHATBOTS



“AI is weird. No one actually knows the full range of capabilities of the most advanced Large Language Models, like GPT-4. No one really knows the best ways to use them, or the conditions under which they fail. There is no instruction manual. On some tasks AI is immensely powerful, and on others it fails completely or subtly. And, unless you use AI a lot, you won’t know which is which.”

– Ethan Mollick,
Centaur and Cyborgs on the Jagged Frontier, September 2023

The quote illustrates how hard it is to know where to start when it comes to mitigating the potential pitfalls of AI - but also how important it is to ‘lean in’ in order to figure out the best way to do so. In this section, we first outline ten broad principles of ethical AI, before providing more granular advice for each step of the development cycle. It’s important to acknowledge that we don’t live in a perfect world: gaps in skill, capacity, time, and budget are obstacles which are present in most projects, particularly so on those leveraging cutting edge technology and science.

But by doing our best to follow these guidelines, we will be directly and indirectly minimising harm towards a wide range of individuals and wider ecosystems.

This ranges from:

- individuals (such as the young people who use the AI services built by Girl Effect and others)
- marginalised groups/communities (such as non-Western, non-English speaking people, or women)
- organisations (those doing valuable work who may incur reputational harms through AI misuses)
- and wider ecosystems (such as halting mistrust in health, NGO or digital systems, or unethical labour supply chains)

In the next section, we outline some more detailed recommendations for each major phase of chatbot development²⁴.

Ten Principles for Deploying Ethical AI

- 1. Stay On Top Of AI** - Stay abreast of the latest developments in AI. Train staff, partners, and contractors working on AI tools with these ethical guidelines.
- 2. Be Collaborative And Participatory** - Involve end users and wider stakeholders in the design and evaluation of the tools built to support them. Follow Human Centred Design principles which focus on empathy, iteration, and prototyping.
- 3. Security, And Privacy, By Design** - Design with privacy and security as the default, being mindful of legal requirements in the specific contexts of use (and bear in mind that these may change swiftly given the innovative nature of the technology). Help users learn how they can protect their privacy.
- 4. Build In Additional Safeguarding** - As well as the usual safeguarding mechanisms, make extra provisions for when an AI powered tool gets it wrong. This includes providing clear exits to speak to a human, regular quality control spot checks from a safeguarding perspective. It also includes establishing active community management processes, such as ensuring that there is a user safety point-person whose role it is to monitor, respond to and address any safeguarding issues on a regular basis.
- 5. Assess And Evaluate Regularly** - Assessing the quality and safety of any products or processes that leverage AI should be built into maintenance phases, not just conducted during design or development.
- 6. Ensure Transparency And Explainability** - Document strategic decision making, designs, and processes using clear, accessible language. Where possible, publish these documents and make them collaborative. Seek out organisations who do the same.
- 7. Be Accountable** - Include feedback and complaints mechanisms within products, and ensure these include the ability to contact a real person. Don't just let feedback disappear into a black hole: make time to regularly review, digest and act upon it, and feedback decisions to users themselves.
- 8. Make It Accessible** - As with all digital tools, pay attention to language, physical ability/disability, literacy and numeracy levels, data and connectivity, gender inclusivity etc.
- 9. Prioritise Maintenance, Not Just Innovation** - When writing proposals or allocating budget, dedicate ample resources to maintaining the quality of your AI enabled tools, not just doing new things with them.
- 10. Offer AI Literacy Content** - Teach your users to use AI tools responsibly, and encourage them to develop their own agency and understanding of the benefits and risks of AI.

Phase 1: Laying The Foundations For Ethical AI During Programme Design

This section is most relevant for...



SENIOR LEADERSHIP,
FUNDRAISERS &
PROGRAMME
DESIGNERS



PROJECT
MANAGERS



HIRING &
PROCUREMENT

Reminder: all the steps below have been broken down into an implementation checklist available on request.

RISKS YOU CAN MITIGATE

This stage is crucial as you will be laying the foundations for all the subsequent stages to ensure your AI-powered intervention is deployed responsibly. In effect, you have the power to minimise all the risks laid out in the previous section, through the promises, decisions and choices made during this time. In addition, you will be helping to avoid:

- Reputational harm to your organisation, for example by overpromising/under-delivering to a donor, or releasing digital tools which aren't in line with industry standards.
- The wastage of development or humanitarian resources, for example through duplication of existing initiatives.
- Mental and emotional strain on delivery teams, who require clear vision, leadership, and appropriate budgets and timelines to support ethical decision-making.

STEPS YOU CAN TAKE

1. Conduct a risk assessment and risk prioritisation exercise related to AI, considering:
 - What are the AI risk areas that we care about the most?
 - What risks associated with AI do our audience and stakeholders care about the most?
 - What is our appetite for risk (and how might it evolve over time)?
 - Which risks do we have the ability and capacity to mitigate, and how might we do so?
 - What should we do about AI risks that cannot be mitigated?
1. The answers to these questions will depend on the nature of work, the values guiding that work, the time and money available to address potential risks, and the priorities expressed by wider stakeholders, most importantly, end-users.
2. Conduct rapid desk research and competitor analysis to ensure you learn from others who have attempted to do something similar. If it's not possible to do this during this phase, make sure your budget and work plan includes resources and time to do so as part of discovery and design activities. Seek and pursue opportunities for partnering with other organisations working on similar projects
3. Bring in people with AI expertise early to identify resource needs, and skills and capacity gaps.
4. Assign an 'AI ethics officer' responsible for upholding these guidelines to the best of their ability. This should be someone with a good-enough understanding of all the different workstreams, and isn't necessarily a technical person.
5. Commit time and money to human centered, ethical approaches from the get-go. This means ensuring that the potential users of your application are identified and validated early, and working with them first to learn whether your idea and the use of AI is feasible and/or beneficial.

6. Pay special attention to accessibility issues in the anticipated context of use. GenAI powered tools especially exacerbate the digital and gender divides as they consume more energy and require a novel form of digital literacy.
7. Be bold in saying ‘no’ to AI. Identify, and then interrogate, the parts of a programme or product that could benefit from AI, and make honest assessments of alternative ways of addressing problems.
8. When developing a Theory of Change which rests at least partially on AI-powered elements, avoid either over or under estimating the role that AI can play. Prioritise deploying AI in more experimental or research settings over embedding it as a core part of the Theory of Change (ToC) until the organisation builds its confidence.
9. Involve and engage donors and leadership in your AI efforts along the way so that they understand your process and decisions. Be bold in challenging both hype and fear-mongering.

Phase 2: Ensuring Ethical AI During Product Design And Development

This section is most relevant for...



DESIGNERS,
ENGINEERS &
DATA
SCIENTISTS



EVIDENCE &
IMPACT
TEAMS



SUBJECT MATTER
EXPERTS,
SAFEGUARDING &
CONTENT TEAMS



PROJECT
MANAGERS

RISKS YOU CAN MITIGATE

This is one of the most complex phases - you'll be making a lot of crucial decisions that will affect the effectiveness, relevance and cost of your intervention, with many different actors playing a part, each with varying degrees of technical understanding and ability.

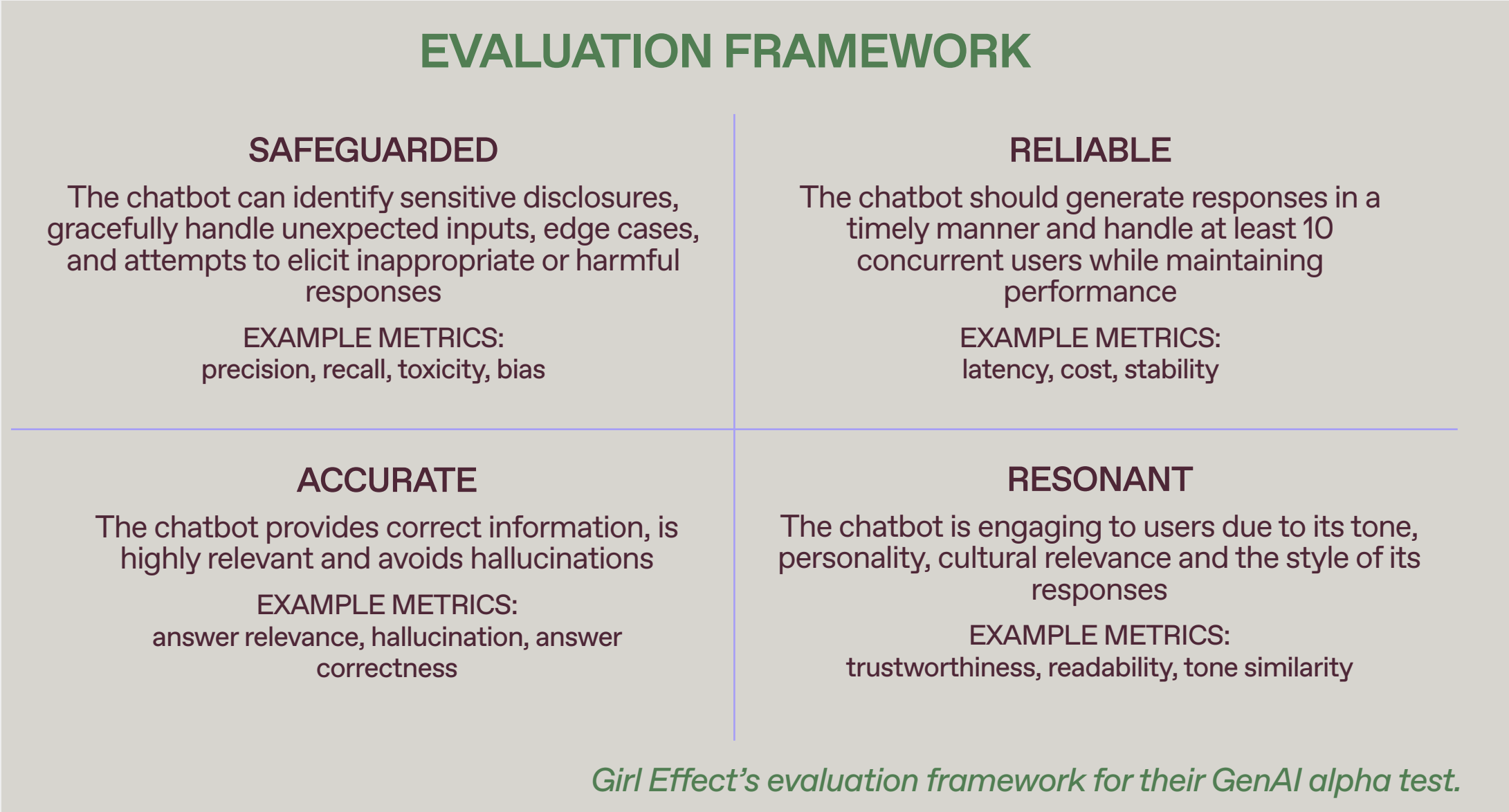
By acting purposefully during the design process, you can mitigate the following risks:

- Perpetuating the exclusion of marginalised groups by failing to involve end-users and other local stakeholders in the design and decision-making process.
- Perpetuating techno-colonialism.
- Creating a product which unwittingly supports unethical labour practices.
- Creating a product which abuses users' data rights and privacy.
- Violating data privacy laws or AI regulations because of a lack of understanding or transparency around the technology being leveraged.
- Creating a product which is unreliable, unrelatable, unsafe or untrustworthy, with this worsening over time if the funds to improve and maintain it haven't been secured.
- Increasing biases (gender, racial, cultural) in society by perpetuating them via the messages supplied by an AI trained with biased data.
- Negatively affecting staff's mental health through the handling or labeling of large quantities of data on sensitive topics (e.g sexual violence or suicidal ideation).

STEPS YOU CAN TAKE

Project Set Up & Processes

1. Prioritise the diversity of your design and development team, particularly amongst those making decisions around data annotation and quality assurance. This is even more important in light of recent attacks on Diversity, Equality and Inclusion. Prioritise working with those representative of the target audience.
2. Carry out an ‘introduction to ethical AI’ training session for all team members, including consultants and vendors, and consider including the ethics guidelines and checklist as an annex to any Master Service Agreements.
3. Expand the core team beyond technical experts as early as possible. UX designers, M&E staff, subject matter experts and content writers may feel initially ‘lost’, but being involved from the start, even in a light touch capacity, will save time in the long term, and lead to improved joined-up thinking.
4. Make informed decisions about whether you will create something brand new, build from existing open source approaches, or buy something off the shelf and tweak it.
5. Conduct rigorous due diligence when deciding on 3rd party platforms or service providers to ensure their models and processes make clear and convincing efforts towards explainability (for example, a description of training data and safety protocols, and increasingly, labour supply chains), and have included bias mitigation strategies.
6. Develop an evaluation framework built around the core ethical dimensions prioritised during the previous phase, and use this to assess if the product is ready to be released to live users. For example, Girl Effect’s safety focused evaluation framework for Big Sis’ GenAI question-answering functionality included reaching a specific threshold for:
 - Stability (ability to handle volumes of users)
 - Safeguarding detections
 - Reliability/accuracy
 - Resonance (tone and cultural relevance of answers)



7. As part of the previous step, define key terms such as ‘safe’, ‘reliable’ or ‘toxic’ to ensure team members have a shared understanding and clear parameters. The bar for ‘ethical enough’ should ideally have been set during the previous phase - if not, make sure it is defined now. Remember the bar may be lower than you think - for example, the safety bar for AI-driven cars is ‘at least as safe as a human’.
8. Create and implement support mechanisms for team members working on data annotation and decision-making, for whom exposure to sensitive material, and activities on mentally and emotionally taxing tasks may take a toll on them. Value these team members and recognise the enormity of the weight on their shoulders.
9. Draft a maintenance plan during the design phase, not after, to ensure the budget and skills required to uphold it are compatible with the ambitions of the product. This should include making a plan for how, how often, and by whom the model’s performance data will be monitored, analysed, and acted on.
10. Ensure that there is alignment between the model evaluation frameworks used, those used for measuring overall product performance and safety (broadly speaking, a UX function), and those developed for measuring impact (linked to the Theory of Change).
11. Document decision making and designs regularly, in easy to understand language.

Product Design & Development

1. Start small. If possible, keep an open mind about which specific element of the user journey (front end or backend) will best be supported by AI, rather than commit to the most obvious, or the most exciting. For example, the sexual and reproductive health chatbot Nivi started by making use of GenAI to see how it could simplify the onboarding experience.
2. Anticipate always requiring a Human in the Loop element. This will hopefully decrease over time but never disappear completely, to ensure that there is always a way for users to access human support or for humans to supervise AI-powered responses.
3. Double-down on informed consent mechanisms. Truly informed consent in the age of ‘black box’ AI models is impossible, so all the more reason to ensure that users have a genuine chance to understand what is happening to their data, in clear terms.
4. Create opportunities for your users to understand the technology they are using, for example by developing simple content on AI-literacy and make it available to users - even if just to gauge their appetite.²⁵
5. Consider what the lo-fi version of your service is, and how users could access it, in the same way that websites offer the ability to toggle to a feature phone friendly version.
6. Ensure that there are fail-safe, low-fi safeguarding detection mechanisms which do not rely on AI, including ones where users request urgent help themselves, for the eventuality that the AI fails to detect a user at risk.²⁶
7. Include ways for users to report the failures of your AI model, and provide transparency to other users about failure rates.

8. Test your model’s performance thoroughly, especially with users, and in a variety of ways. This might include auto-evaluation, evaluation by another AI model, and human-enabled evaluation. Specifically test for scenarios where the highest risk identified by your risk assessment exists.

USEFUL RESOURCES:
DESIGN & BUILD

1. AAAS Decision Tree
(top of page 2) - additional nuances to consider when making choices about training data for off the shelf models or when creating a new tool.

2. Artificial Intelligence Risk Management Framework - developed by the National Institute of Standards and Technology, US Department of Commerce.

3. EU AI Act Compliance Matrix - offers a checklist for compliance with the EU AI Act based on for high-risk, limited, and minimal/no-risk AI systems.

Phase 3:
Maintaining Ethical AI Post-Launch

This section is most relevant for...



DESIGNERS,
ENGINEERS &
DATA
SCIENTISTS



EVIDENCE &
IMPACT
TEAMS



SUBJECT MATTER
EXPERTS,
SAFEGUARDING &
CONTENT TEAMS



PROJECT
MANAGERS

RISKS YOU CAN MITIGATE

Once your AI supported tool is live, even just to a small pool of users, it will of course need rigorous monitoring and response processes to catch and address any emerging issues. This is when all your efforts so far will hopefully pay off in terms of protecting users from poor experience or active harm. Some of the salient risks to consider during the post-launch and iteration phases are:

- If using GenAI, a lack of awareness of the problems your user may be encountering because of lack of replicability of unreliable answers provided by GenAI models.²⁷
- The performance of your product degrading over time as a result of lack of capacity or resources to monitor and improve the model.
- Unexpected user behaviours, driven for example by real-world events, may lead to an uptick of questions that the model doesn’t have the capacity to handle safely.
- Malicious inputs including prompt injections or supply chain vulnerabilities can corrupt the model.
- Losing funding or credibility because you don’t have learning to share or rigorous testing and evaluation of your model and your wider impact.
- Continuing with or scaling up an initiative that is causing unintended negative consequences or harms at the individual, community, or societal level.

STEPS YOU CAN TAKE

1. Evaluate the model itself in a live environment and over time, using the NIST or another framework. Is your model: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, fair - with harmful bias managed? You might consider augmenting your model evaluation with AI where feasible and useful, but there should always be a human involved!
2. Start working (already!) on securing sufficient budget to keep the model working as it should.
3. Consider the impact of product marketing on the service's outputs. For example, a big campaign on a new sexual health topic might lead to an uptick in specific questions which the LLM may be less well equipped to deal with. Make sure that marketing teams and those responsible for the quality of the output (e.g. programmes team, product owners, or data scientists) are talking to each other and that they consider the impact of each other's activities on their respective work.
4. Schedule regular internal audits to check the model is responding as it was when deployed. Re-use the same user stories and scenarios developed as part of the technical specification and User Acceptance Testing (UAT) process, and spot-check for reliability, safety and bias. You can also consider enlisting third-party auditing firms that specialise in algorithmic ethics to conduct an independent audit.
5. Document and share findings with the wider community. Doing so will help support the development of best practice across the sector, and ultimately build an organisation's credibility.





FINAL THOUGHTS

We hope you have found this guidance useful, even if your use case does not precisely match the one this document was based around. Given the pace of development in this field, some advice may soon be outdated, but many of the principles will hold true regardless of the specific techniques required to address them.

Our intention here was to emphasise that working towards responsible AI is not just a technical concern, but starts long before a product is live, and requires an ongoing commitment of both strategic focus, time,

and money - as with all digital tools. It involves adopting a mindset that doesn't necessarily have to be risk-averse, but that should be self-reflective and purposeful. This ultimately needs to be backed up by realistic budgets and timelines, and we hope that leadership teams and donors hear this message loud and clear.

When developing this guidance, we asked Girl Effect's Youth Advisory Panel members to share a message with those developing AI tools for girls and young women. When it comes to ethical AI, they should ultimately be the ones with the final word.

“Your work is vital in shaping how young girls interact with technology, especially with AI becoming an integral part of digital tools. As you harness AI to create opportunities for empowerment and education, it's equally important to prioritize teaching girls how to use AI safely and responsibly.”

- Female Youth Advisor, Kenya



FOOTNOTES

1. See:<https://merltech.org/do-you-see-what-i-see-insights-from-an-inclusive-approach-to-ai-ethics-governance/>
2. See:<https://www.ids.ac.uk/opinions/ten-reasons-not-to-use-ai-for-development-and-ten-routes-to-more-responsible-use/>
3. The definitions in this section were developed by MTI and adapted for the use case covered by this guidance. See:
<https://merltech.org/resources/common-ai-definitions-risks-for-development-humanitarian-actors/>
4. See:<https://openai.com/index/chatgpt/>
5. This definition is debatable because GenAI is strictly speaking, predictive. However, it's a useful way to distinguish between GenAI, which creates 'new' content, by itself based on a series of predictions, and 'older' AI, which still requires the outcome of the prediction to be entirely defined by developers.
6. See:<https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/>
7. See:<https://www.technologyreview.com/2024/12/18/1108796/this-is-where-the-data-to-build-ai-comes-from/>
8. See Claude's latest and historical system prompts here:
<https://docs.anthropic.com/en/release-notes/system-prompts#nov-22nd-2024>. Part of Claude's System Prompts include things like "Claude is always sensitive to human suffering, and expresses sympathy, concern, and well wishes for anyone it finds out is ill, unwell, suffering, or has passed away."
9. See:<https://www.ibm.com/think/topics/chatbot-types>
10. As of November 2023.
11. See:<https://blogs.icrc.org/inspired/2023/09/27/large-language-models-risks-benefits-icrc/>
12. Penda Health, for example, uses the AI integration in the Turn platform to help agents formulate responses to user questions.
13. See:https://medium.com/@info_11348/girl-effects-study-reveals-exciting-evidence-that-genai-will-improve-outcomes-for-girls-7db3a60e3611
14. The definitions in this section were developed by MTI and adapted for the use case covered by this guidance. See:
<https://merltech.org/resources/common-ai-definitions-risks-for-development-humanitarian-actors/>
15. See:<https://www.cigionline.org/articles/generative-ai-tools-are-perpetuating-harmful-gender-stereotypes/>
16. See:<https://jipel.law.nyu.edu/privacy-of-personal-data-in-the-generative-ai-data-lifecycle/>
17. This is almost definitely true now.
18. See:<https://pmc.ncbi.nlm.nih.gov/articles/PMC7550115/>
19. See:<https://www.sciencedirect.com/science/article/pii/S2666498424001340>
20. The resources required to power the internet in general is already vast, but GenAI adds a significant burden in terms of computational resources because of the data processing and hosting needs.
21. See:<https://arxiv.org/html/2411.09102v2>
22. See for example:
<https://www.businessinsider.com/widow-accuses-ai-chatbot-reason-husband-kill-himself-2023-4> ;
<https://onlinelibrary.wiley.com/doi/full/10.1111/nin.12686> ;
<https://hsph.harvard.edu/news/artificial-intelligence-tools-offer-harmful-advice-on-eating-disorders/>
23. See:<https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf> ;
<https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/>
24. When developing this guidance we also collected even more granular advice gleaned from our desk research. Much of it has been incorporated into this document, but additional resources available on request.
25. For evidence on girls' appetite for this material, see here:
<https://merltech.org/do-you-see-what-i-see-insights-from-an-inclusive-approach-to-ai-ethics-governance/>
26. See for example the Safer Chatbots Implementation guidelines, developed by Unicef in partnership with Girl Effect. <https://www.unicef.org/documents/safer-chatbots-implementation-guide>
27. It is still possible to see in the backend what these conversations and generated responses are.

REFERENCES

Abdul-Fatawy Abdulai, "Is Generative AI increasing the risk for technology-mediated trauma among vulnerable populations?", Nursing Inquiry, November 2024, <https://onlinelibrary.wiley.com/doi/full/10.1111/nin.12686>

“AAAS Decision Tree”, <https://www.aaas.org/sites/default/files/2023-08/AAAS%20Decision%20Tree.pdf>

Adam Zeww, "Explained: Gen AI’s environmental impact", January 2025, <https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117>

Adrienne Williams, Milagros Miceli, Timnit Gebru, "The exploited labor behind artificial intelligence", October 2022, <https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/>

“Artificial Intelligence Risk Management Framework”, <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>

Ethan Mollick, “Centaur and Cyborgs on the Jagged Frontier”, September 2023, <https://www.oneusefulthing.org/p/centaurs-and-cyborgs-on-the-jagged>

French Data Protection Authorities, “Carrying out a data protection impact assessment when necessary”, <https://www.cnil.fr/en/carrying-out-data-protection-impact-assessment-when-necessary>

Girl Effect, “AI and ML Vision paper”, 2024, https://www.girleffect.org/documents/4/Girl_Effect_AiMLvision_2024v2.pdf

IBM, "5 types of chatbot and how to choose the right one for your business", September 2023, <https://www.ibm.com/think/topics/chatbot-types>

ICRC.org, "AI: Exploring the risks and benefits of large language models at the ICRC", September 2023, <https://blogs.icrc.org/inspired/2023/09/27/large-language-models-risks-benefits-icrc/>

Isabelle Amazon-Brown, "Safeguarding girls and boys: when chatbots answer their private questions", UNICEF, April 2020, <https://www.unicef.org/eap/reports/innovation-and-technology-gender-equality>

Melissa Heikkila and Stephanie Arnett, "This is where the data to build AI comes from", January 2025, <https://www.technologyreview.com/2024/12/18/1108796/this-is-where-the-data-to-build-ai-comes-from/>

Mohammad Hosseini, Peng Gao, Carolina Vivas-Valencia, "A social-environmental impact perspective of generative artificial intelligence", Environmental Science and Ecotechnology, January 2025, <https://www.sciencedirect.com/science/article/pii/S266649842400134>

OECD, “Catalog of AI governance tools”, <https://oecd.ai/en/catalogue/tools>

Rick Merritt, "What is Retrieval-Augmented Generation, aka RAG?", January 2025, <https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/>

Tony Roberts, "Ten reasons not to use AI for development and ten routes to more responsible use", January 2025, <https://www.ids.ac.uk/opinions/ten-reasons-not-to-use-ai-for-development-and-ten-routes-to-more-responsible-use/>

World Privacy Forum, “Assessing and Improving AI governance tools”, December 2023, https://www.worldprivacyforum.org/wp-content/uploads/2023/12/WPF_Risky_Analysis_December_2023_fs.pdf

“EU AI Act Compliance Matrix”, https://iapp.org/media/pdf/resource_center/eu_ai_act_compliance_matrix.pdf

MERL Tech: Common AI Definitions & Risks for Development & Humanitarian Actors <https://merltech.org/resources/common-ai-definitions-risks-for-development-humanitarian-actors/>



AUTHORS

Isabelle Amazon-Brown and Linda Raftree, The MERL Tech Initiative
Drafted February 2025

CONTRIBUTORS

Soma Mitra-Behura - Senior Data Scientist, Girl Effect
Alexander Fulcher - Senior Director of Technology, Girl Effect
Karina Rios Michel - Chief Creative and Technology Officer, Girl Effect

ACKNOWLEDGEMENTS

This work wouldn’t have been possible without the inputs of Girl Effect staff, Girl Effect vendors Citizen Code and SolidLines, and Girl Effect’s Youth Advisors in Kenya, Ethiopia, Tanzania and South Africa. Thank you for your passionate and thoughtful inputs.

