safer chatbots

implementation guide

unicef for every child
table of contents

Background 02

Who is this guide for? 03

The end product 03

Requirements & skills 04

Step 1: Assess your current safeguarding response 05

Step 2: Identify useful stakeholders 06

Step 3: Conduct a co-creation workshop 07

Step 4: Identify referral services 08

Step 5: Finalise keywords & responses 09

Step 6: Implement technical changes 10

Monitoring & maintenance 12

Useful resources 13

Contact 14

Acknowledgements 14
The Covid-19 pandemic led to an urgent need for the rapid dissemination of trustworthy health information at an unprecedented scale. This resulted in a proliferation of messaging-based services, or chatbots, which provide information to hard-to-reach audiences in a conversational format, via channels such as SMS, WhatsApp and Facebook Messenger. These chatbots sometimes include a level of Artificial Intelligence (AI) to interpret users’ messages, but more often are built as pre-defined ‘decision trees’ with a conversational feel. Whilst these services have played a powerful role in communicating vital health messaging, users who may be experiencing intense hardship often see these chatbots as a way to seek help with potentially life-threatening issues.

In particular, Gender-based violence (GBV) increased as a result of the mental and economic pressures placed upon individuals and families, and the social distancing measures imposed by governments during the pandemic. This put vulnerable women and children at increased risk of violence, made worse by disruption to GBV, MHPSS and CP services. In these cases, digital channels were often the only recourse for women and children at risk of further harm.

However, chatbots built rapidly in response to the pandemic are mostly built without sophisticated AI, therefore unable to recognise many user inputs or respond appropriately. Most also do not include an option to speak to a human (expert, counsellor, case manager). In addition, most ‘error’ responses do not cater to urgent help-seeking behaviours. This failure to respond appropriately can put users at increased risk of harm. This is not unique to Covid-19 chatbots: a Learning Brief¹ on chatbots for sexuality education and adolescent reproductive health conducted by UNICEF EAPRO in February 2020 had already highlighted this significant safeguarding issue, showing that most chatbots are designed without provision to detect and respond to children and adolescents at risk of serious harm.

The Safer Chatbots project intends to address this situation by providing tried and tested blueprints for the inclusion of improving safeguarding measures into any chatbots reaching users who may be at risk, especially girls and women, as well as children in vulnerable contexts, such as refugees and migrants.

who is this guide for?

Safer Chatbots offers 3 different mechanisms, depending on the level of technical sophistication of your chatbot, as outlined in the Safer Chatbots explainer video. This guide covers the second highest level of sophistication, providing high-risk keyword based recognition and response. Keyword-based recognition is, of course, an imperfect solution, lacking the ability to detect more nuanced disclosures of harm, but we believe it is better than no detection at all. The approach in this guide has been piloted by UNICEF and WHO chatbots in 3 countries - U-Report Tanzania and Uganda, and HealthBuddy Bulgaria, as well as in external chatbots such as Big Sis by Girl Effect, and Vibe Check by the Prince’s Trust International.

This guide was developed for any chatbot creator who uses RapidPro or Textit to communicate with their users via SMS, Facebook Messenger, Telegram or WhatsApp (although the approach could easily be adapted for other chatbot platforms). Your chatbot does not need a Natural Language Processing (NLP) engine, although you can still benefit from these guidelines if you have. You do not need to be an engineer or AI expert to use this guide, but a decent understanding of RapidPro/Textit by at least one team member is important.

the end product

By the time you’ve implemented the Safer Chatbots mechanism, when users send your chatbot a message which includes one or more words indicative of a disclosure such as violence or self-harm, they will receive an automated response checking if they need to talk to someone. If they confirm, they’ll be provided with a reassuring and compassionately worded message including key information on appropriate services they can self-refer to. We recommend you try out the chatbots of our three pilot countries to see the mechanism in action (See: Useful Resources).
to get there, you will need to achieve 5 things:

1. A list of common trigger words used by your users when disclosing abuse or risk of harm.
2. A global safety word to communicate to users at the start of the chatbot experience.
3. Messages to use in the automated response to user disclosures.
4. A list of vetted service providers to refer users to.
5. A list of the places in your chatbot where the safeguarding mechanism needs to be included.

requirements

- A RapidPro or textit account.
- A messaging based service, either live or planned.
- Access to the documentation listed under Useful Resources.

team skills

- Good knowledge of RapidPro/Textit.
- Good knowledge of your user’s contexts and habits.
- Good at writing content for your target audience.
- Ability to identify and consult with experts in Child Protection, Gender Based Violence or similar.
step 1 assess your current safeguarding response

Look at your users’ conversations with your chatbot, and start collating examples of disclosures (for example: “I was violated”, “I want to report a case of child abuse”) as well as how your chatbot responds to them. What themes emerge? What words come up frequently? Are they misspelled? (For example, in one of our pilot countries, a user shared “my little sister is being forsd to marry, please help”, highlighting the need to include the common misspelling for ‘forced’.) You should always consider where in your chatbot these disclosures have happened. How could your chatbot have done better in its response?

Now take a first look at the generic list of trigger words (based on the work done by the Safer Chatbots pilot countries, see: Useful Resources) What thematic areas are missing? What words are missing? (for example, if trafficking or financial violence is more prevalent in your context, you may want to ensure the list covers that area). It’s particularly important to seek out words relevant to your particular context, for example, the word ‘defiled’ is used more commonly in Uganda than the term ‘rape’.

Finally, take a look at the generic response messages (see: Useful Resources), as well as the localised versions used by the pilot countries. What changes might you make for your own context?

thanks to this exercise, you should now have:

😊 A list of sentences and words your users have historically used to disclose a case of abuse or self-harm
😊 A list of thematic areas and words used by other countries for comparison
😊 A draft list of flows where improvements need to be made
😊 Ideas for how your chatbot’s response should improve

N.B: If you don’t have a chatbot yet, you can still do this exercise, just skip the review of user conversations in favour of a discussion of how you anticipate your users might communicate, and what provisions, if any, you’ve planned for detecting and responding to users in distress.
The Safer Chatbots pilot teams all worked with internal and external experts to develop the final list of trigger words as well as the automated safeguarding responses. We recommend reaching out to individuals or groups with experience in Gender Based Violence, Child Protection, Protection against Sexual Exploitation and Abuse, and Mental Health, and inviting them to be part of an in person or remote consultation.

Finally, make sure that you have identified a group of user advisors, who might be existing users of your service or representative of your target audience, and who can be involved in the development of trigger words and the automated responses. Establish the best communication channels for them - whether it’s email, a Teams meeting, or a WhatsApp group.

"During the pandemic, refugee and migrant girls and women are especially vulnerable to violence. This is made worse by the closure of refugee and GBV support centres. We partnered with service providers with deep knowledge of refugee and migrant needs to craft culturally sensitive responses to those using our digital channels to get help, as well as referral information for services provided in their mother tongue."
Once you’ve identified your stakeholders, organise a workshop to get their inputs. Here’s a sample agenda from our Safer Chatbots project lead in Tanzania:

- **Introduce participants** to the Safer Chatbots initiative using assets in the Useful resources section.

- **Demo** the Safer Chatbots mechanism using one of the pilot countries’ chatbots. If possible, demo how your chatbot currently responds to highlight the changes needed.

- **Ask for participant inputs** on the importance and relevance of this project.

- **Present your 1st draft** of keywords; split into teams and challenge them to propose additions, removals or changes. Get them to pay particular attention to text speak, misspellings etc. Reconvene, discuss and collate inputs.

- **Present the generic response** to a possible user disclosure; split into teams and challenge them to propose changes. Teams present their feedback as part of a role play activity. Vote for the best one, or co-write a final version based on everyone’s ideas.

- **Brainstorm a single emergency word** to allow people to jump straight to safeguarding information in the future (for example: SAUTI was used in Tanzania).

- **Wrap up** - agree on commitment to feedback on the final mechanism once it’s been developed.

“We engaged with youth from various walks of life, including refugees and those with disabilities. They gave us constructive feedback during the workshops, which we’ve incorporated to our safeguarding flows to ensure that we speak with young people in a way we’ll be heard.”

Mariam Mmbaga
U-Report Consultant | Tanzania
This is one of the most important steps because we have a duty to ensure that all users are referred to services that will meet their needs and not put them at increased risk of harm. To help you agree on the number and nature of the services you will signpost users to, ask yourselves the following questions:

- Do they provide services appropriate to the type(s) of disclosure we are likely to receive via the chatbot?
- Have they been vetted by Child Protection colleagues or similar?
- When we contact the service as a user would, do we get the expected response? (for example, hotline open at advertised time, response to an email...)
- Will the user be helped directly, or will their case be passed on to another service (for example, a local government based service), requiring potentially painful repeated disclosures, and possible confidentiality issues?
- Most importantly - can we commit to regularly monitoring and updating this referral information? The answer to this question should influence the number of referral services you choose to include.

“...it was tempting to include a long list of all the available referral services in the safeguarding response. But signposted services need to be sustainable options, with the best chance of long-term availability to beneficiaries, and that we could feasibly commit to a vetting on a regular basis. We therefore prioritised national, well-established services who could in turn connect youth and children to local service providers.”

Fatuma Namukose
U-Report | Uganda
Thanks to your workshop, you should now be in a position to finalise your list of trigger words, the global emergency keyword, as well as the messages included in your automated response. Take your time on this step, if necessary, returning to your stakeholders to get their final sign off. In particular, consult your user-advisors and Child Protection and Safeguarding colleagues to get their sign off.

In particular, do a final review asking the following questions:

😊 Are there any common misspellings we’ve not included?

😢 Are any of these words likely to be used in a different context not indicative of harm? If so, remove them.

😄 Is the global keyword easy to type and to remember, and unlikely to be used in normal conversation?

Once your keywords are finalised, use a tool like this one² to turn the columns into a comma separated list. You’ll need to use this format when adding them to RapidPro/TextIt.

²https://convert.town/column-to-comma-separated-list
The diagram below shows how your system should respond to disclosures. Most importantly, given that keyword-based detection is an imperfect solution, you should always check with the user whether they were indeed trying to seek help, allowing them to continue their experience if not. You’ll find two .json files included in the Useful resources section which will show you how the trigger words get added to RapidPro flows, and how users get passed back and forth from the Safeguarding flow (where the automated compassionate responses and referral information lives).
Trigger words should be added at each ‘Wait for Response’ branch using the ‘has any of the words...’ function, and categorised as ‘Safeguarding’. Any messages categorised as Safeguarding should send the user to the Safeguarding flow. Once that flow is completed by the user (either because they indicate they don’t actually want human assistance or because they’ve completed the Safeguarding flow) they should be sent to the point they left off in the original flow. This creates a more seamless, discrete experience for users, minimising disruption to their interaction with your service. There’s no getting away from the fact that this step requires a bit of repetitive work! You could alternatively add these as individual trigger words in the Triggers section, however each will need to be copy > pasted individually. Doing so at flow level also incurs a limit on the amount of triggers you can include. Or, if you have the support of a developer, you could connect via webhook to a cloud database of the keywords.

Finally, set up your global safety word in the Triggers tab. This will allow users to jump straight to the Safeguarding flow at any time in the future. Make sure to add a new message to your onboarding experience so that users are informed about this trigger word from the start. A sample registration flow including some messaging you might want to use, including around privacy, is included in the project documentation.

Make sure you thoroughly test the new mechanism using as many of the trigger words as you can, in as many different flows. Challenge your colleagues and user advisors to try it out and ‘break’ the system so you can identify any bugs or gaps early.
how to monitor & maintain the safer chatbots mechanism

A detailed set of recommendations for monitoring and maintenance is included under Useful resources. The most important indicators to track are:

😊 How frequently your users are triggering the Safeguarding flow.
😊 What they typed to trigger it.
😊 Whether they’re choosing to continue.
😊 If yes, whether they complete the flow.

To create a collection of users who have accessed the safeguarding flow, you might want to add them to a group at the start and end of the flow. You can then browse the conversation history of users in that group to understand more about any changes you might need to make.
useful resources

The documents listed below are included in the Useful Resources folder which accompanies this guidelines document.

1. To try U-Report Tanzania, message SAJILI to 15070 (SMS), or to +255746039550 (WhatsApp) or via m.me/ureport.tz/ (Facebook Messenger)

2. To try U-Report Uganda, message JOIN to 8500 via SMS

3. Safer Chatbots explainer video

4. General tips on improving the safety of chatbots: Safeguarding girls & boys: when chatbots answer their private questions', UNICEF EAPRO Learning Brief, April 2020

5. Generic list of trigger words

6. Generic automated safeguarding response

7. Sample registration flow flagging the emergency safety word and showing integration of trigger words into flow ‘wait for response’ (json file)

8. Sample safeguarding flow (json file)

9. Safer Chatbots: Monitoring & Maintenance advice

10. To use the Turn Playbook: Safeguarding templates, contact Pippa@turn.io

11. To learn about, adapt or use the NLP model, contact Girl Effect (karina.michel@girleffect.org) and Weni (yves.bastos@weni.ai)
contact us

The Safer Chatbots project was developed by UNICEF East Asia Pacific, Gender section. We hope the guidance provided has been easy to follow, and will help you on your journey to making every chatbot safer for women, girls, and vulnerable individuals.

If you have any questions, or would like to know more about the project, including the two other options for improving the safety of your chatbots, contact Isabelle Amazon-Brown and Gerda Binder via gbinder@unicef.org.

acknowledgements

The Safer Chatbots project would not have been possible without the generous support of the US Bureau for Population, Refugees and Migration. Special thanks also to UNICEF HQ - Child Protection, and to our colleagues from the three pilot countries in Tanzania, Uganda and Bulgaria Country Offices, as well as to all the service providers and youth advisors who supported them. Finally to our external partners, the teams at Girl Effect, Weni, Chayn and Turn.