safer chat bots

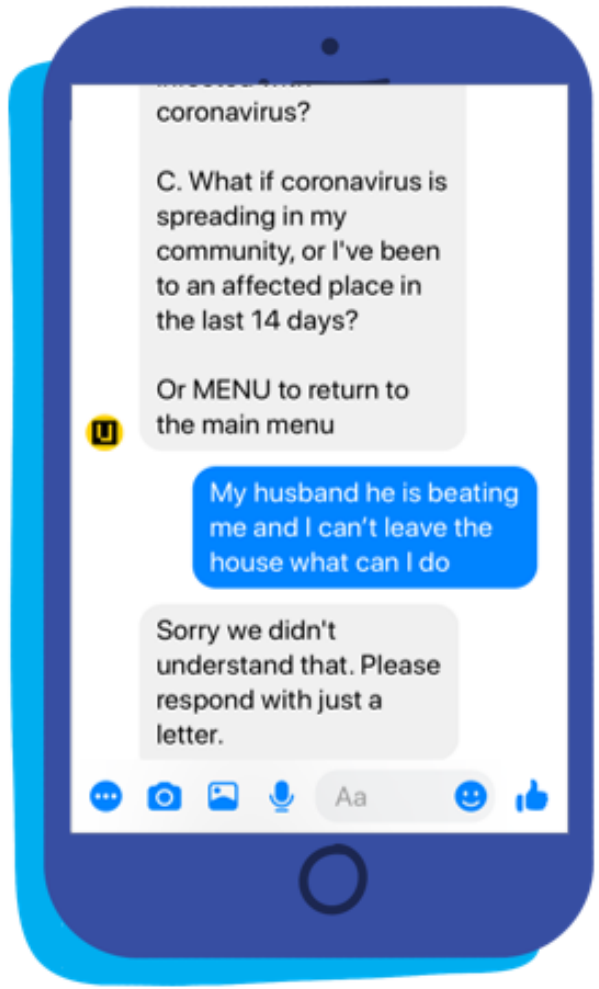
what are chatbots?

Chatbots are messaging-based services which provide information to audiences in conversational format, via channels such as SMS, WhatsApp or Facebook Messenger.

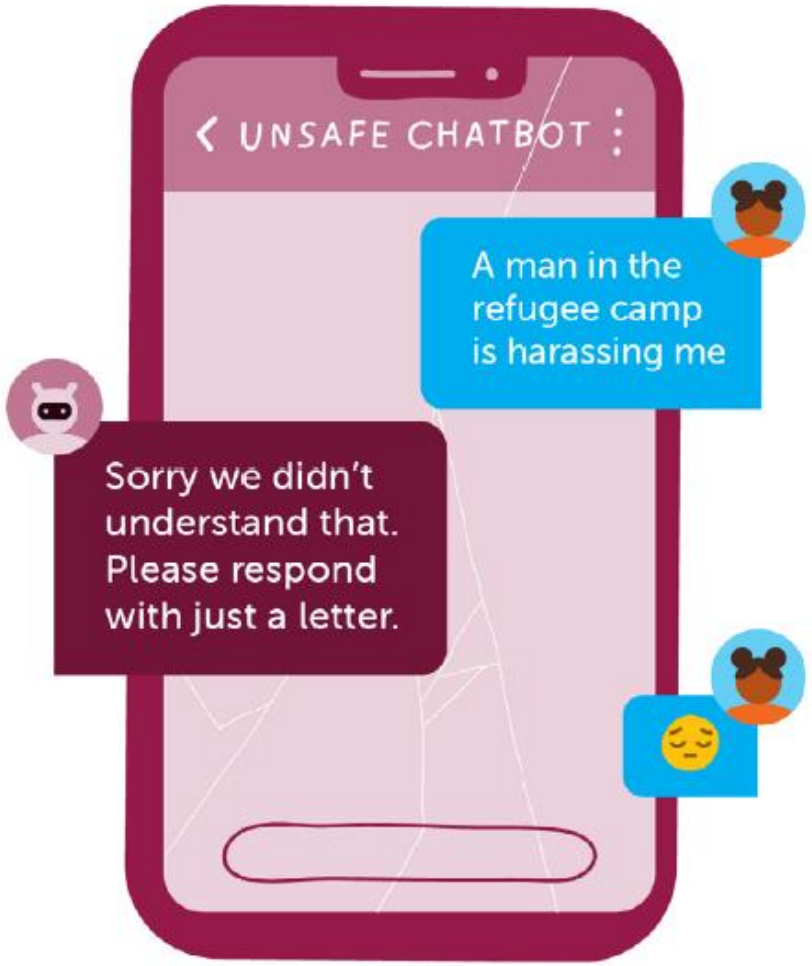
Chatbots sometimes include a level of Artificial Intelligence (AI) to interpret users' messages, but more often are built as pre-defined 'decision trees' with a conversational feel.

why safer chatbots?

Many chatbots, including those powered by AI, don't do a good enough job at detecting and responding to users disclosing traumatic or life threatening situations such Gender Based Violence or suicidal ideation.



piloting safer chatbots



Together with COs (Uganda, Tanzania, Bulgaria) and partners (Chayn, GirlEffect, Weni, Turn) the Safer Chatbot project developed and illustrated mechanisms to include safeguarding flows for users disclosing GBV or seeking help.

These mechanisms range from a single emergency word, to keyword-based detection, to an AI model.

safer chatbots everywhere - implementation guidance

1

Step 1: How does your chatbot currently respond to users in distress?

How does your chatbot currently respond to users in distress? What words do your users tend to use to describe their situation? How could you do better?

2

Step 2: Identify useful stakeholders

What expertise do you lack when it comes to identifying useful keywords and a context appropriate response to users in distress?

3

Step 3: Co-creation workshop

Work with your useful stakeholders, including young people, to validate keywords and responses & create something truly relevant for your users.

4

Step 4: Identify referral services

Make sure the services you refer users to are reliable, sustainable, and relevant to users' needs.

5

Step 5: Finalise keywords & responses

Ensure your chosen keywords and response messages cover things like mispellings and colloquial expressions by testing a prototype with stakeholders.

6

Step 6: Technical implementation

Add your keywords to relevant flows in RapidPro or similar, connect them to the compassionate responses & referral information, and ensure your users can smoothly continue their conversation after a safeguarding side-chat.

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.

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 developed by Girl Effect and Weni, contact Weni via weni.ai

resources

1

Explainer video to help introduce your team to the project

2

Learning brief to provide background on the issue of chatbots and safeguarding

3

Overview document to give your team a quick sense of what implementation involves

4

Detailed implementation guidelines for keyword based safeguarding detection

5

Open access NLP model if you're working with AI