6 steps to make your chatbot safer for children & young people

Young people and children often use chatbots to get help with traumatic or life-threatening situations - even when these chatbots aren't designed to provide such support. Most chatbots haven't been set up to detect and respond to users in distress, increasing the risk of further harm. The Safer Chatbots project has created 3 different mechanisms to help you increase the safety of your chatbot, depending on its level of sophistication and the resources available to you. The steps below are useful for those whose chatbot is not AI powered, and who have used RapidPro or similar to build their chatbot. It focuses on programming the detection of individual keywords indicative of a high-risk situation, the deployment of a global 'safe word' users can type at any time, and a series of compassionate response messages, along with details for reliable referral services.

unicef ♥ | for every child

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?

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?

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.

Step 4: Identify referral services

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

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.

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.

For full technical guidelines, refer to the Safer Chatbots Implementation guide. Safer Chatbots is an initiative by UNICEF East Asia Pacific, Gender section. If you would like know more, including about the two other options for improving the safety of your chatbots, contact Isabelle Amazon-Brown and Gerda Binder via gbinder@unicef.org.