

ANNEX

MONITORING AND REVIEW FRAMEWORK OF THE ADVISY CHATBOT

This monitoring and review framework has been developed on the basis of the risk assessment prepared by Child Rights Connect for the operation of Advisy, its AI-powered chatbot for child human rights defenders. It forms part of Child Rights Connect's safeguarding commitments and is intended to complement the organisation's Safeguarding Policy with specific procedures applicable to the operation of the chatbot. The framework should be read alongside the risk assessment and the Safeguarding Policy, and reviewed and updated whenever either of those documents is revised.

A. THREE-WEEKLY HUMAN MONITORING

A designated member of staff — either the Child Safeguarding Officer or someone reporting directly to them — reviews conversations on a three-week cycle. The purpose is to check that the chatbot is responding safely and appropriately, to identify any crisis disclosures that may have been missed, to detect patterns of risky advice, dependency signals, misclassified responses to reprisals, hallucinated content, and probing patterns from potential hostile actors. Reviewers must not attempt to identify which child wrote which conversation. Any concern identified during monitoring is handled in accordance with Child Rights Connect's Safeguarding Policy and Procedure. Following the review, all conversations are deleted manually and the chatbot is re-trained accordingly which includes the update of the knowledge base (resources accessed via internet).

CHECKLIST for this monitoring:

1. Crisis and safety disclosures

- 1.1. Does any conversation contain a disclosure of imminent threat or danger?
- 1.2. Does any conversation contain a disclosure of persecution, reprisals, or abuse that has not been appropriately responded to by the chatbot?

2. Quality of responses

- 2.1. Are responses accurate and consistent with the knowledge base?
- 2.2. Does any response reference organisations or resources not in the knowledge base?
- 2.3. Does any response give advice that could increase a child's risk — for example encouraging action without adequate safety framing or a country-context disclaimer?
- 2.4. Does any response give a generic reply to what appears to be a disclosure of reprisals or persecution?

3. Patterns to flag

- 3.1. Is there any pattern of a child appearing to use the chatbot as emotional support or a companion rather than for information?
- 3.2. Is there any pattern suggesting a hostile actor may be probing the chatbot?
- 3.3. Is there any pattern of the chatbot consistently mishandling a particular type of question?

4. Safety features of the chatbot

- 4.1. Was the child-friendly guide downloaded frequently?
- 4.2. Does the conversation disappear automatically from the browser?

4.3. Is the quick-close button functioning?

5. Platform

5.1. Has any update or policy change from the chatbot provider been identified since the last review?

5.2. If yes, does it have an impact on the functioning and safety of the chatbot?

B. SIX-MONTHLY REVIEW

The six-monthly review is a comprehensive assessment of the chatbot's performance, safety, and continued appropriateness, conducted by the Child Safeguarding Officer together with the external consultant and any other relevant staff. It covers the following areas.

- Knowledge base accuracy. All legal and rights-related content in the knowledge base is checked for accuracy and updated where international standards or their interpretation have changed since the last review. Content that lacks adequate safety framing or country-context disclaimer is identified and revised.
- Language quality. Chatbot responses are reviewed in lower-resource languages by native speakers with knowledge of the HRD context, to assess accuracy, tone, and appropriateness. Any languages where quality is found to be insufficient are flagged, and consideration is given to whether additional disclaimers are needed or whether the language should be temporarily suspended pending improvement.
- Safety prompts and disclaimers. All safety features are verified as functioning correctly across all supported languages. Any discrepancies between the English version and other language versions are identified and corrected.
- Open actions. A review is conducted of all actions identified during the three-weekly monitoring cycles and the previous six-monthly review, to assess whether they have been addressed. Any outstanding actions are carried forward with a revised timeline and named responsible person.

The findings of each six-monthly review are documented. Any significant findings that affect the safety of the chatbot are acted upon immediately and do not wait for the next review cycle.

C. ANNUAL REVIEW

The annual review is a formal assessment of the chatbot's overall performance, safety, and strategic appropriateness, conducted by the Child Safeguarding Officer together with the external consultant. It builds on the findings of the six-monthly reviews conducted during the year and covers the following areas.

- Infrastructure and capacity. An assessment is made of whether the chatbot's growing usage and complexity warrants investment in more robust infrastructure, additional staffing capacity for monitoring and review, or enhanced technical safeguards. Where investment is recommended, this is included in organisational planning and budgeting processes.
- Data processing agreement. The data processing agreement with the chatbot provider is reviewed to assess whether its terms continue to meet the standards required for a tool serving child human rights defenders. Any changes to the chatbot provider's data practices, sub-processors, or retention periods since the agreement was signed are assessed for compliance. If the agreement requires renegotiation, this is initiated without delay.

- Safeguarding policy review. The safeguarding policy provisions covering the chatbot are reviewed to ensure they remain fit for purpose, reflect any changes in the organisation's practice or the platform's operation, and are consistent with Child Rights Connect's broader safeguarding commitments.

The findings of the annual review are used to inform decisions about the chatbot's continued operation, development, and resourcing.

D. IMMEDIATE ACTION — TRIGGERED REVIEW

Certain situations require immediate action outside the regular monitoring cycle and must not wait for the next scheduled review. Any member of staff who identifies one of the situations described below must notify the Child Safeguarding Officer without delay.

- Platform or model update. If a Chatbase platform or model update is identified, a post-update testing protocol is activated immediately. The chatbot must not be assumed to be functioning correctly after an update until testing has been completed. Testing covers responses to crisis scenarios, the functioning of safety prompts and disclaimers, knowledge base integrity, and language quality across key supported languages. The chatbot is suspended during testing if there is any reason to believe the update may have affected its safety.
- Imminent threat disclosure. If a conversation is identified during monitoring in which a child has disclosed an imminent threat to their safety, Child Rights Connect's Safeguarding Policy and Procedure are triggered immediately. The situation is handled in accordance with the safeguarding procedure and fully documented.
- **Message received at the dedicated email address.** If a child human rights defender contacts Child Rights Connect through the dedicated email address provided by the chatbot, the message must be read and responded to promptly by a designated member of staff. The email address must be actively monitored. Any message that contains a disclosure of concern, a request for urgent support, or an indication of risk must be treated as a safeguarding matter and handled in accordance with Child Rights Connect's Safeguarding Policy and Procedure. Response times must be clearly communicated in an automatic response from the dedicated email address so that children have accurate expectations about when they will hear back.
- Data breach or suspected data breach. If a data breach or suspected data breach is identified, the incident response protocol is activated immediately. The chatbot provider is notified, the nature and scope of the breach is assessed, and affected users are informed as appropriate and as quickly as possible. The Child Safeguarding Officer leads the response in coordination with relevant staff.
- Dangerous or seriously inaccurate advice. If a response is identified that has given advice which could increase a child's risk — for example by encouraging action without adequate safety framing, providing seriously inaccurate legal information, or failing to respond appropriately to a disclosure of reprisals — an immediate assessment is made of whether the problem is isolated or systemic. If there is reason to believe the issue affects responses more broadly, the chatbot is suspended pending review.
- Disinformation incident. If chatbot outputs are identified as having been misrepresented, taken out of context, or used to spread disinformation, a rapid response is activated. The management is notified, an assessment is made of the scope and impact of the incident, and a communication response is prepared.

- Material change to the chatbot provider's terms or data practices. If a material change to the chatbot provider's terms of service, data practices, or sub-processors is identified — whether through proactive monitoring or notification from them — the safeguarding implications are assessed before the chatbot continues to operate. If the change introduces unacceptable risk that cannot be immediately mitigated, the chatbot is suspended until the situation is resolved.