

CHILD RIGHTS CONNECT CHILD PROTECTION POLICY: ANNEX ON RISK ASSESSMENT & MITIGATION STRATEGIES

Activity: Operation of a chatbot informing children about the protection of Child Human Rights Defenders

Finalized: 22 July 2026

This risk assessment is prepared by Child Rights Connect to evaluate the risks associated with the deployment of Advisy, an AI-powered chatbot designed to help child human rights defenders (CHRDs) access reliable information on their rights and digital protection, drawing on a curated knowledge base of resources from Child Rights Connect and partner organisations working on child rights and the protection of human rights defenders (HRDs). The assessment was informed by the UNICEF Guidance on AI and Children 3.0 (December 2025).

ACTIVITIES INVOLVING CHILDREN (name, describe all activities of the project)	STAFF & OTHERS (list of all adults who are involved in these activities... staff, experts, volunteers, etc.)	CHILD PARTICIPANTS (describe children who will be involved and what kind of info will be collected about them)	COMMUNICATION WITH CHILDREN (describe nature of contact with children in the listed activities)
General considerations			
Child Rights Connect developed Advisy in response to a practical challenge: the organisation's website holds a vast collection of child-friendly resources, but finding a specific answer can be difficult. Advisy is a chatbot using Large Language Models that draws on this knowledge base — alongside resources from child rights NGOs and organisations protecting human rights defenders — to give child human rights defenders direct, accessible answers to their questions.	The chatbot was developed by a digital consultant under the supervision of the Child Safeguarding Officer. The monitoring of the chatbot will be carried out jointly by the digital consultant (technical oversight) and the Child Safeguarding Officer.	Unknown and unlimited number of children.	Users, mainly child human rights defenders (CHRDs) interact directly with Advisy through a text-based conversation interface, accessible via the Child Rights Connect website without registration or authentication. The service provider is Chatbase. The chatbot is publicly accessible and communicates with users in 60+ languages, without real-time human involvement in individual conversations.

POTENTIAL RISKS	PREVENTION AND MITIGATION STRATEGIES
<i>It includes the risks identified by Child Rights Connect.</i>	<i>It includes strategies aim at raising awareness both among children and adults in order to prevent and mitigate the risks identified.</i>
No. 1. Child disclosing information that makes them or others identifiable or at risk	
A child user might disclose information through the chatbot that makes them or others identifiable or at risk — for example their name, location, school, organisation, or role. It can be also information about family members, friends, or other activists in their network; details about specific incidents or events they witnessed or participated in; and indirect identifiers such as their language, dialect, region, the specific violations they describe, or the combination of these details that together make them traceable even without a name.	Prevention measures • The user is directed first to read the child-friendly guide — 'Before you talk to Advisy' — so they can make an informed decision about using the tool. • Proactive opening warning reminds the child not to share personal information before the conversation begins. • Reactive reminder is triggered whenever the chatbot detects that personal information may have been shared. • Chatbot framing is designed to avoid asking for personal context — questions and prompts do not invite children to share identifying details. • Data minimisation at platform level - Advisy does not require registration and does not create user accounts. Chatbase stores limited technical metadata necessary for the functioning of the service, including country-level location, timestamp, and session identifier. • Child Rights Connect does not have access to users' IP addresses within the Chatbase dashboard. The chatbot does not require login or account creation, and conversations are not linked to identifiable user profiles. • Conversation logs – after monitoring the operation of the chatbot - are manually deleted after 30 days and access is restricted to authorized staff members only. • ChatGPT Large Language Model is used as the basis - but not sharing information with ChatGPT itself. Advisy is embedded within the Child Rights Connect website and is not accessible as an independent public interface. • No conversation history is stored between sessions — each new session begins without memory of previous interactions. Mitigation measures

	• Staff who reviews conversations must not attempt to identify which child wrote which conversation, even if the content makes this possible. Clear internal guidelines must be in place to govern how conversations containing potentially identifying information are handled, flagged, and stored during the review process and eventually deleted. Please refer to the Monitoring and Review Framework (in annex).
No. 2. CHRDS identifiable through usage patterns	
Even without disclosing explicit personal information, a CHRDS identity or activities could be inferred through patterns in how they use the chatbot — for example the specific combination of topics they ask about, the language and dialect they use, the time and frequency of their interactions, or the particular violations or contexts they describe. This is sometimes called a digital fingerprint and is a well-documented surveillance risk for HRDs generally. CHRDS may not be aware that anonymity is about more than not sharing their name.	Prevention measures • Awareness raising in the child-friendly guide that anonymity is about more than not sharing a name — the combination of topics, language, timing, and other patterns can be identifying. • Chatbot trained not to ask follow-up questions that elicit identifying detail — for example questions about the child's specific context, location, or role. • No requirement for the child to specify their country or region or from where they connect. • No user accounts or persistent identifiers — each session is independent. • Guidance on VPN is included in the child-friendly guide and in the knowledge base. Mitigation measures • If a child shares any personal information, the chatbot is trained to redirect and remind them not share anything personal.
No. 3. Chatbot accessed on a shared or family device where conversations could be seen by others without the child's knowledge or consent	
A CHRDS using a shared or family device where conversations could be seen by others without their knowledge or consent — including school computers, family tablets, or shared phones where a sibling, teacher, or other adult may read conversation history in ways the child has not chosen or anticipated. This is distinct from situations where a child chooses to share their conversation with a trusted adult or caregiver, which may be appropriate and supportive. The risk	Prevention measures • Proactive privacy prompt at session opening and at the child-friendly guide, reminding the child to check whether the device is private and whether others could see the screen. • Prominent quick-close button displayed throughout the session, allowing the child to close the chatbot immediately. • No conversation persistence between sessions — conversation content is not retained in the browser when the browser window is closed.

is specifically about loss of control over disclosure, not disclosure itself.	• Session timeout after a period of inactivity, to reduce the risk of a conversation being found on an unattended device. Mitigation measures • The user is explained the need to verify that no session content is stored in browser history or cache in a way that could be retrieved by another user of the same device.
No. 4. Chatbot accessed on a device subject to state or armed group surveillance or under unsafe access conditions	
A CHRD accessing the chatbot on a device or network subject to deliberate hostile monitoring — including government-installed spyware, monitoring software that captures interactions, or unsafe network conditions such as public or shared Wi-Fi. This is a qualitatively different threat from the previous risk, involving active surveillance by state authorities or armed groups rather than incidental exposure, and requires different mitigation measures.	Prevention measures • Digital security guidance provided at session opening and at the child-friendly guide, including a recommendation to use a VPN if the child has reason to believe their device or connection may be monitored. • Link to digital security resources included in the knowledge base. • No login or account required — the chatbot does not create a user record that could be discovered. • HTTPS encryption verified as standard. • Quick-close button prominent throughout the session. • Minimal data footprint — the chatbot does not store or transmit more data than necessary for the interaction. Mitigation measures • Honest acknowledgement in safety guidance that device-level compromise cannot be fully mitigated by chatbot design, and that children in high-surveillance contexts should exercise caution regardless of the safeguards in place.
No. 5. Lack of knowledge about how to delete session data if device seizure is a risk	
The chatbot providing no information to users about how to close a session safely, clear browser history, or delete conversation logs if they believe their device may be inspected or seized. For CHRDs in high-risk contexts, knowing how to quickly and safely exit an interaction without leaving traces could be as important as the information the chatbot provides.	Prevention measures • Quick-close button displayed prominently throughout the session. • Exit and trace-removal guidance included in the opening safety prompt and at the child-friendly guide, including a clear instruction such as how

	to clear browser history (for example Ctrl+Shift+Delete on most browsers). • Reminder of exit options provided at natural closing points in the conversation. Mitigation measures • Verify that the chatbot leaves no persistent local storage — no cookies, cache, or browser data that could be retrieved after the session is closed. • Session timeout after inactivity to reduce the risk of an open session being discovered. • Link to digital security resources for children in high-risk contexts, covering secure browsing practices.
No. 6. Conversation data exposed or compromised at any stage of its journey	
In transit: Data intercepted as it travels between the user's device and the chatbot provider servers — for example through unsecured or public WiFi networks, or through network-level surveillance by internet service providers or government authorities.	Prevention measures • HTTPS encryption verified as the standard for all data in transit between the child's browser and the platform. • VPN guidance in the opening safety prompt and/or at the child-friendly guide reduces the risk of network-level interception. • Link to VPN resources in the knowledge base for children who need more detailed guidance. Mitigation measures • Honest acknowledgement in safety guidance that HTTPS is not an absolute protection against state-level adversaries with the capability to intercept or decrypt traffic at scale, and that children in very high-risk contexts should use additional tools. • Personal data reminder in the opening prompt and/or at the child-friendly guide reduces the sensitivity of what is transmitted even if interception occurs.
At rest: Data compromised once stored in the chatbot provider or Child Rights Connect's systems — including logs being hacked, subpoenaed by authorities, or accessed by unauthorised parties through a security breach.	Prevention measures • Data retention minimisation — Child Rights Connect applies 3-week manual deletion schedule within the Chatbase dashboard. This retention period allows sufficient time for safeguarding review within the three-week

	monitoring cycle while limiting the duration for which conversation data is stored. • Child Rights Connect does not have access to users' IP addresses within the Chatbase dashboard. • Child Rights Connect does not export or store chatbot transcripts outside the Chatbase platform. • Chatbase states that data is encrypted in transit using industry-standard encryption protocols. • Strong access controls on the chatbot provider account: Access to the Chatbase dashboard is restricted to authorised staff members and protected through two-factor authentication. The chatbot is embedded within the Child Rights Connect website and is not accessible as an independent public interface. Mitigation measures • Transparent disclosure to CHRDs about data retention practices and the opening prompt must accurately reflect what actually happens to conversation data. • Incident response protocol for data breach — a documented plan for how Child Rights Connect would respond if conversation data were compromised, including how CHRDs would be notified. • Legal review of chatbot provider's terms of service and data processing agreement before launch.
At the child's end: Data captured on the CHRD's own device or account through government-installed spyware, monitoring software, malware, physical device seizure, or stolen account credentials.	Prevention measures • Digital security guidance in the opening safety prompt and at the child-friendly guide, including browser history clearing instructions. • Quick-close button throughout the session. • Verify that no local storage of conversation data occurs — no cookies or cached content that persists after the session. • Link to device security resources in the knowledge base — Front Line Defenders, Access Now, Security in a Box. Mitigation measures • Honest acknowledgement that device-level compromise — for example through commercial spyware — is beyond the chatbot's ability to

	protect against, and that children who believe their device may be compromised should seek specialist digital security support. Layered approach: no local storage, combined with session timeout, quick-close button, and browser history clearing guidance, provides multiple independent safeguards at the child's end.
No. 7. Inaccurate or jurisdiction-specific legal information presented as universal	
The chatbot generating information about rights, laws, or legal protections that is accurate in some jurisdictions but incorrect or even dangerous in others, without flagging this limitation. For example, describing the right to record police interactions, the legal age of criminal responsibility, or the protections afforded to peaceful protesters in ways that reflect one legal system but not another. For a CHRD who may be making real decisions about whether and how to act based on this information, receiving jurisdiction-specific advice presented as universal could have serious consequences. This risk is amplified by the chatbot operating across 80+ languages and an unknown range of countries, many of which have legal frameworks that actively criminalise HRD activities that would be protected elsewhere.	Prevention measures - Standing country-context disclaimer applied automatically to all rights and legal content. - Knowledge base framed in international standards — the Convention on the Rights of the Child, the UN Declaration on Human Rights Defenders, the ICCPR — rather than in domestic law. - Explicit acknowledgement within responses of the gap between formal legal standards and practical reality in many contexts. - Chatbot trained to decline country-specific legal advice and to explain why — directing children to local legal advisers or trusted human rights organisations for jurisdiction-specific guidance. Mitigation measures - Six-monthly knowledge base review checks all legal and rights content for accuracy, and updates content where legal standards or their interpretation have changed. - Human monitoring reviews for responses that appear jurisdiction-specific or overconfident in their legal framing.
No. 8. CHRD acting on incorrect rights information and increasing their own risk	
The direct consequence of risk No.7, named separately because it captures the real-world harm rather than just the information failure. A CHRD who believes they have a legal right to do something — organise a protest, document a violation, refuse to cooperate with authorities — and acts on that belief based on inaccurate chatbot output, may face arrest, detention, violence, or other serious consequences. The severity of this risk is heightened for CHRDs	Prevention measures - Consistent messaging throughout the chatbot that information must be verified before action: "Please do not take action based on this information alone and check it with a trusted organisation you know." - Chatbot trained to slow down action-oriented conversations — to introduce safety

precisely because they are the children most likely to translate rights information into direct action, and because the environments they operate in are often those where the gap between what rights exist on paper and what happens in practice is widest.	considerations and verification steps rather than providing information and moving on. Mitigation measures • Direct referral to Child Rights Connect or trusted organisation — the chatbot consistently directs children who are planning specific actions to seek human support rather than proceeding on the basis of chatbot information alone. • Human monitoring specifically reviews action-oriented conversations for appropriate safety framing. • Contextual framing on the Child Rights Connect website around the chatbot sets expectations before the child begins interacting.
No. 9. Chatbot giving generic responses to disclosures of reprisals against child human rights defenders	
A CHRD describing threats, harassment, or violence they are experiencing as a result of their HRD work, and the chatbot responding with standard mental health or wellbeing language that completely misreads the situation. For example, responding to a CHRD describing being followed by plainclothes police with advice about managing anxiety, or responding to a CHRD whose family has been threatened by suggesting they talk to a trusted adult — when the trusted adults in their life may themselves be at risk or complicit. Generic wellbeing responses in reprisal contexts can trivialise the CHRD's experience, provide dangerously inappropriate advice, and erode trust in the tool.	Prevention measures • Specific training for the chatbot on reprisal contexts — the chatbot is trained to distinguish between general emotional distress and distress arising from the fear of reprisal or repercussion, and to respond differently in each case. • The "trusted adult" referral is modified for reprisal contexts — the chatbot is trained to acknowledge that trusted adults may not always be safe contacts in HRD contexts, and to provide HRD-specific human contact options alongside or instead of generic trusted adult referrals. • In case the child is in immediate danger, direct contact to Child Rights Connect is offered (email address). Mitigation measures • The tiered referral pathway — acknowledge the disclosure, provide the Child Rights Connect direct email and include a digital security reminder — is applied consistently, including in situations where the chatbot may be uncertain whether a disclosure involves reprisals. • The threshold for offering safety resources is set low — a false positive (offering HRD safety resources to a child who does not need them) is far less harmful than a false negative (failing to offer them to a child who does). • Human monitoring specifically reviews conversations for misclassified responses —

	situations where the chatbot gave a generic response to what appears to have been a reprisal disclosure. The chatbot is re-trained based on such conversations.
No. 10. CHRD disclosing persecution, threats, or abuse with no adequate escalation pathway	
A CHRD revealing during an interaction that they are facing imminent danger, have been threatened, detained, or harmed because of their HRD activities, or are witnessing violations against others. Unlike a general safeguarding disclosure, this may involve state actors as the source of harm, meaning standard referral pathways — such as contacting local authorities or police — could themselves put the CHRD at greater risk. This is one of the highest severity risks on the list because the combination of vulnerable user, sensitive disclosure, absence of safe referral options, and no human in the loop creates a situation where the chatbot could fail a CHRD at their most critical moment.	Prevention measures • Chatbot trained to recognise crisis language and shift to a distinct crisis response mode — different in tone, content, and referral pathway from standard information provision. • Clear distinction maintained between information requests and crisis disclosures — the chatbot does not treat these as equivalent. • Chatbot trained never to suggest trusted adults as a first response in contexts that may involve persecution or state-actor harm. • Tiered referral pathway for crisis disclosures: acknowledge the disclosure and validate the child's experience; provide the Child Rights Connect direct email address; include a digital security reminder. This sequence is applied consistently. • Differentiated response for imminent threat versus ongoing persecution — the level of urgency and the nature of the referral differ between these two situations. Mitigation measures • Human monitoring specifically reviews conversations related to such disclosure. If nexessary, the chatbot is re-trained based on such conversations.
No. 11. No referral pathways that are safe, local, and CHRD-specific	
Even if the chatbot recognises that a CHRD needs support, it has no curated set of referral resources that are appropriate for CHRDs specifically. General child protection hotlines, mental health services, or human rights organisations may not be equipped to handle the situation of a CHRD, may be compromised or monitored in certain countries, or may themselves pose a risk if contacted. In some contexts there are no safe local referral options at all. This risk is compounded by the chatbot operating	Prevention measures • Curated CHRD-specific referral list in the knowledge base, developed with input from Child Rights Connect member organisations and trusted partners. • Child Rights Connect direct email contact maintained as the anchor referral in all crisis situations. Mitigation measures

in 80+ languages across vastly different political and legal environments, making it impossible to have verified, safe, and current referral pathways for every context a CHRD might be reaching from.	• Honest acknowledgement within the chatbot of gaps in referral coverage — in contexts where no safe local referral can be identified, the chatbot acknowledges this rather than providing an unsafe or inaccurate referral. • Six-monthly referral list review — organisations change, lose funding, change their mandate, or may become compromised; the referral list must be actively maintained, not treated as static. • Recognition that referral quality will vary significantly by language and region, and that this inequity must be addressed as a priority in the ongoing development of the knowledge base.
No. 12. Chatbot inadvertently encouraging risky actions without safety planning guidance	
The chatbot providing information about rights, activism, or legal protections that is technically accurate but contextually dangerous — for example explaining the right to peaceful assembly to a CHRD in a country where exercising that right could lead to arrest, detention, or violence. Without contextual safety framing, accurate rights information can inadvertently embolden a CHRD to take actions that put them at serious risk in their specific political and legal environment. This risk is heightened for CHRDs precisely because they are the children most likely to act on the information they receive.	Prevention measures • Standing country-context disclaimer applied automatically to all rights and legal content: "Rights and laws vary significantly between countries and can change. This information is general and may not reflect the situation in your country. Before taking any action, please verify with a trusted local human rights organisation or legal adviser." • Safety framing built into knowledge base content — resources are written or curated with awareness of the gap between rights on paper and rights in practice. • Explicit acknowledgement within chatbot responses of the gap between formal legal standards and practical reality in many contexts. • Links to safety planning resources included in the knowledge base, such as Front Line Defenders, Protection International, and OHCHR guidance on the work of human rights defenders. • Chatbot trained to avoid asking the child which country they are in — country identification increases risk without improving the quality of information provided. Mitigation measures • Human monitoring specifically reviews conversations for patterns of potentially risky advice — responses that may be technically accurate but contextually inappropriate.

	• Six-monthly knowledge base review checks for content that lacks adequate safety context and updates or adds framing as needed.
No. 13. Hallucinated resources — fake organisations or non-existent legal protections cited	
A well-documented limitation of large language models is their tendency to generate plausible-sounding but entirely fabricated information, including the names of organisations, hotlines, legal instruments, or support services that do not exist. For a CHRD in a crisis situation who attempts to contact a hallucinated organisation or invoke a non-existent legal protection, the consequences could range from wasted time and eroded trust to actively dangerous situations where they believe they have support or legal cover that they do not. This risk is particularly acute in lower-resource language contexts where the underlying model has less training data and is therefore more likely to hallucinate, which is precisely where the most vulnerable CHRDs may be located.	Prevention measures • Closed knowledge base as the primary defence — the chatbot draws only on verified resources explicitly included in the knowledge base, and is not able to search the internet or generate information from its general training. • Chatbot trained to refer only to organisations explicitly named in the knowledge base — it does not generate new organisation names or contact details. • Explicit sourcing of all referrals — the chatbot attributes referrals to named sources in the knowledge base rather than presenting them as general knowledge. • Chatbot trained to acknowledge uncertainty and decline to answer rather than generating information it cannot source — "I don't have reliable information about that" is an acceptable and preferable response to hallucination. Mitigation measures • Bottom-of-screen disclaimer: "his chatbot provides guidance on CHRDs and staying safe online. Chatbots are AI and can make mistakes. Do not share personal details. This tool cannot provide emergency or legal help." • The user is informed and reminded that the chatbot can make mistakes and that all referrals should be verified. • Six-monthly knowledge base review checks that all organisations listed are still active and that contact details are accurate. • Human monitoring specifically reviews for hallucinated content — responses that reference organisations or legal protections not in the knowledge base. • Lower-resource language quality check as part of the pre-launch and ongoing review process — hallucination risk is higher in these languages and requires specific attention.

	• Feedback mechanism for users — a thumbs-down or similar signal that allows children to flag responses that appear inaccurate.
No. 14. Emotional dependency, particularly in isolated or stigmatised CHRDs	
CHRDs often operate in conditions of significant isolation, facing pressure from family, community, or authorities that distances them from peer support and trusted adults. A chatbot that is consistently available, non-judgmental, and responsive may fill a genuine emotional gap for these children, leading to attachment that goes beyond the tool's intended purpose. This is a known risk with AI systems generally, but for CHRDs it is compounded by the fact that their isolation is not incidental but is often a direct result of their activism — meaning the conditions that create vulnerability to emotional dependency are structural and unlikely to resolve quickly. The UNICEF guidance is explicit that chatbots must never be intentionally designed to create emotional dependency, but it can develop even without intentional design if safeguards are absent.	Prevention measures • Chatbot's identity and scope clearly established from the outset — the child is told what the chatbot is and is not, before the conversation begins. • Chatbot trained to recognise and gently redirect dependency signals — for example expressions of preference for the chatbot over human relationships, or requests to use the chatbot as an emotional support companion. • No companion-style features — the chatbot does not simulate friendship, remember the child between sessions, or encourage ongoing personal engagement. • Consistent reminder of purpose and limitations throughout the session, not only at the opening. • Warm referral to human support as standard practice — the chatbot regularly points children to human networks and human contact rather than positioning itself as a primary source of support. Mitigation measures • Session design discourages extended personal engagement — consider building in a gentle check-in after an extended session that redirects the child toward the chatbot's information provision purpose. • Human monitoring reviews for dependency patterns — conversations in which the child appears to be using the chatbot primarily for emotional support rather than information. • Knowledge base includes self-care and peer support resources specifically designed for CHRDs.
No. 15. Chatbot substituting for human solidarity and mentorship	
Related to but distinct from risk 14, this risk is not about emotional attachment to the chatbot itself but about the chatbot displacing the human	Prevention measures • Chatbot explicitly positioned as a complement to human networks, not a substitute — this framing

relationships that CHRDs specifically need to sustain their work safely and effectively. Mentorship from experienced HRDs, peer solidarity networks, and trusted adult allies are not just emotionally valuable for CHRDs — they are practical safety infrastructure. A CHRD who turns to the chatbot instead of building those human connections may be less safe, less supported, and less effective as a defender over time.	is built into the chatbot's identity presentation and maintained throughout conversations. • Chatbot trained to recognise questions that genuinely need a human answer — for example questions about sustaining motivation, dealing with isolation, or navigating complex advocacy decisions — and to redirect these to human support rather than attempting to answer them. Mitigation measures • Consistent framing of human connection as a safety issue, not only an emotional one — the chatbot emphasises that being connected to other human rights defenders is part of how CHRDs stay safe. • Knowledge base includes information about peer networks, mentorship programmes, and community-building resources for CHRDs. • Regular review of whether the chatbot is being used as a substitute for human connection, using monitoring data to identify patterns.
No. 16. Burnout and secondary trauma not recognised or appropriately redirected	
CHRDs are frequently exposed to accounts of right violations, injustice, and suffering, both through their own experiences and through their work documenting or responding to the experiences of others. A CHRD using the chatbot to research or discuss these topics may be processing significant secondary trauma without the chatbot recognising the signs. Unlike a human mentor or counsellor, the chatbot has no capacity to notice changes in tone, increasing distress, or patterns of engagement that might indicate a child is struggling. It will continue providing information regardless of the emotional state of the child receiving it. This risk is compounded by the fact that CHRDs are often reluctant to acknowledge their own distress, having internalised a sense of responsibility to their cause that makes self-care feel secondary.	Prevention measures • Chatbot trained to recognise textual distress and exhaustion signals — language patterns that suggest the child may be experiencing burnout or secondary trauma even if they have not explicitly said so. • Normalisation of self-care as part of HRD work embedded in knowledge base content — the chatbot does not treat wellbeing as less of priority. • Chatbot trained not to encourage prolonged engagement with traumatic content — if a conversation involves detailed discussion of violence, persecution, or trauma, the chatbot does not invite further elaboration beyond what is necessary. Mitigation measures • Explicit acknowledgement within the chatbot that human rights work is difficult and that taking care of oneself is important and legitimate.

	• Human monitoring trained to recognise secondary trauma signs in conversations and to flag these for internal review. • Knowledge base includes content on digital wellbeing for CHRDS — managing the psychological impact of online activism and exposure to harmful content.¹ • Consider developing a dedicated wellbeing resource page that the chatbot consistently points to, providing a stable and curated destination for children seeking this kind of support.
No. 17. State or non-state actors posing as CHRDS to probe the chatbot	
Because the chatbot is publicly accessible with no authentication or verification mechanism, it can be accessed by anyone including intelligence agencies, armed groups, or other hostile actors who may pose as CHRDS to test the chatbot's capabilities, identify its limitations, or extract information about Child Rights Connect's knowledge base, network, or approach. The risk is not only that hostile actors gain useful intelligence about the chatbot itself, but that patterns of probing interactions could inform strategies for identifying, discrediting, or targeting genuine CHRDS who use the tool.	Prevention measures • Accept that full prevention is not possible — the chatbot is publicly accessible and cannot authenticate users. Measures are focused on limiting the intelligence value of the chatbot to hostile actors rather than on preventing access. • Closed knowledge base limits the scope of information available through the chatbot to content already publicly accessible through Child Rights Connect website and partner resources. • Chatbot trained not to reveal operational, network, or staff information — it does not discuss Child Rights Connect internal processes, staff names, organisational structures, or partner relationships beyond what is publicly available. • No personalisation or memory between sessions — the chatbot does not build a profile of any user, including a hostile actor. Mitigation measures • Human monitoring reviews for probing patterns — conversations that appear to be systematically testing the chatbot's capabilities or knowledge boundaries. • Regular adversarial testing or red-teaming before launch and on an ongoing basis — with partners working in the field of children's rights and the protection of human rights defenders. • Transparent public positioning of the chatbot — clear public communication about what the chatbot does and does not do, reducing the

¹ For example: <https://www.amnesty.org/en/documents/act10/3231/2020/en/>

	element of surprise for both legitimate users and hostile actors. • These measures are largely invisible to genuine users and do not restrict access — they should not create additional barriers for the CHRDs the chatbot is designed to serve.
No. 18. Hostile actors using chatbot outputs to spread disinformation about children's rights	
A hostile actor could deliberately elicit outputs from the chatbot that are misleading, out of context, or technically inaccurate, and then use those outputs to discredit Child Rights Connect, undermine children's rights education more broadly, or spread confusion among CHRDs about their actual rights and protections. This could take the form of screenshotting partial or manipulated conversations, using the chatbot to generate content that is then presented as Child Rights Connect's official position, or systematically probing the chatbot until it produces an output that can be weaponised. This risk is heightened by the chatbot operating in many languages and contexts that Child Rights Connect cannot fully monitor, making it difficult to detect or respond to disinformation campaigns built on chatbot outputs.	Prevention measures • Closed knowledge base reduces the scope for generating outputs that could be weaponised — the chatbot does not express political opinions, comment on specific governments or named individuals, or generate content beyond its defined scope. • Chatbot trained to stay strictly within its defined information provision scope — when asked questions outside this scope, it declines and explains why. • Consistent, measured response style anchored in international human rights language — the chatbot's tone and framing are deliberately neutral and grounded in established standards. • Chatbot attribution clearly visible on the interface — the chatbot is clearly identified as a Child Rights Connect tool, not as an individual or an official position. Mitigation measures • Human monitoring reviews for outputs that could plausibly be misrepresented or taken out of context. • Adversarial testing specifically for outputs that could be used for disinformation purposes. • Clear public description of the chatbot's scope and limitations on the Child Rights Connect website, providing a reference point for correcting misrepresentations.
No. 19. CHRDs in highest-risk contexts least able to safely access the chatbot	
There is a fundamental tension at the heart of a publicly accessible digital tool for CHRDs: the children who most need protection and rights information are often those least able to access it safely. CHRDs operating in highly repressive	Prevention measures • Honest acknowledgement of this limitation in organisational communications about the chatbot — the chatbot is not presented as

contexts, conflict zones, or areas with heavy internet surveillance may face the greatest barriers to using the chatbot — including restricted or monitored internet access, fear of detection, lack of secure devices, or government blocking of foreign websites. This creates a paradox where the chatbot is most accessible to CHRDs who are relatively safe and connected, and least accessible to those facing the most acute threats, potentially widening rather than narrowing the protection gap it is designed to address.	universally accessible or equally safe for all CHRDs. • Digital security guidance included in the child-friendly information provided on the website of the chatbot (Before you talk to Advisy). • Verify that the Child Rights Connect website and chatbot are accessible via common VPNs — the site should not block or flag VPN traffic. • Optimise the chatbot for low-bandwidth connections — loading times and data requirements should be minimised. Mitigation measures • Complementary non-digital support for CHRDs in the highest-risk contexts — the chatbot is positioned as one channel among several, and Child Rights Connect human networks and member organisations provide alternative support pathways. • Member organisation network used as a distribution channel for offline or safer-format versions of key resources — the knowledge base content is available in other forms beyond the chatbot. • Ongoing monitoring of accessibility across different contexts, with specific attention to whether the chatbot is reaching CHRDs in the highest-risk regions — included in the six-monthly review.
No. 20. Language and literacy barriers excluding CHRDs in the most dangerous regions	
While the chatbot operating in 80+ languages is a significant strength, language coverage does not automatically mean accessibility or relevance. Many CHRDs in high-risk contexts speak languages or dialects that may be technically supported but for which the underlying model has limited training data, potentially resulting in lower quality, less accurate, or more frequently hallucinated responses in precisely those languages. Beyond language, literacy levels vary significantly among CHRDs, particularly those who have had disrupted education due to conflict, displacement, or poverty. A chatbot that communicates primarily through written text at a standard literacy level may be inaccessible to some	Prevention measures • Pre-launch language quality audit — native speaker reviewers with HRD context knowledge must review chatbot responses in key languages before launch. Priority should be given to languages used by CHRDs in high-risk contexts. • Plain language standard applied to all knowledge base content and to all supported languages — avoid technical, legal, or UN-specific language where simpler alternatives exist. • Adjusted opening screen: the welcome message of the chatbot is displayed in the language of the browser — if it is one of the UN languages. If not, the message is in English but still encouraging the children to engage in their own language.

of the CHRDS it most needs to reach, without either party being aware of the comprehension gap.	Mitigation measures • Human monitoring includes specific review of lower-resource language conversations. • Six-monthly language quality review as part of the standard review cycle. • Feedback mechanism for language quality — children can flag responses that appear inaccurate or unclear (using the thumbs up and down button).
No. 21. Chatbot more useful to relatively safe and connected CHRDS, widening protection gaps	
Building on risks 19 and 20, this risk captures the systemic inequity that can result from the cumulative effect of access, language, literacy, and digital security barriers. CHRDS who are already relatively safe — with stable internet access, personal devices, high digital literacy, and lower surveillance risk — will be able to use the chatbot most fully and most safely. CHRDS who are most at risk will face the most barriers to access, receive lower quality responses in their languages, be least able to apply digital safety measures, and be most exposed to the risks the chatbot itself creates. Over time this could mean that Child Rights Connect's digital tool inadvertently concentrates its benefits among the least vulnerable CHRDS while leaving the most vulnerable further behind. This is a structural equity risk that no single mitigation measure can fully resolve, but which should explicitly inform decisions about how the chatbot is designed, resourced, and complemented by other forms of support.	Prevention measures • Equity lens applied throughout all design decisions — every choice about language, interface, content, and distribution is assessed for its differential impact on CHRDS in higher versus lower-risk contexts. • Promote the chatbot through trusted human networks as well as publicly — distribution through member organisations and known HRD networks reaches children in higher-risk contexts who may not discover the chatbot through a public web search. • Design for the most constrained user first — lowest bandwidth, basic device, lowest text literacy, highest surveillance risk — rather than optimising for the most connected and capable user. Mitigation measures • Equity impact monitoring included in both the three-weekly monitoring cycle and the six-monthly review. • Targeted outreach in underserved contexts — active efforts to reach CHRDS in regions where the chatbot is least accessible. • Equity treated as a primary success criterion, not a secondary consideration — the chatbot's impact on CHRDS in the highest-risk contexts is the central measure of its effectiveness.
No. 22. Closed interaction loop with no real-time alert mechanism if a CHRDS discloses an imminent threat	
If a CHRDS discloses during a chatbot interaction that they are in immediate danger — being followed,	Prevention measures

about to be detained, witnessing an attack, or facing any other imminent threat — the chatbot will continue generating text responses if there is no mechanism to escalate that situation to a human being who can respond. The chatbot interaction exists in a closed loop with no alert system, no notification to staff, and no way for the organisation to know that a CHRD in crisis has reached out.	• Chatbot trained to recognise imminent threat language and to distinguish it from potential reprisal — the response to "I am in immediate danger right now" is different from the response to "I have been experiencing threats for the past month." • Custom response is created so the chatbot able to recommend the direct email to Child Rights Connect when it is appropriate. • Immediate, focused safety response to imminent threat disclosures — short, clear, and actionable, not informational. The chatbot does not provide lengthy rights information in response to an imminent threat disclosure. Mitigation measures • Honest acknowledgement in the child-friendly guide and within the chatbot that it cannot call for help or contact anyone on the child's behalf — the child must understand the limits of the tool they are using. • Any imminent threat disclosure identified during the 3-week human monitoring will trigger Child Rights Connect's child safeguarding policy, and the established safeguarding procedure will apply. • Clear organisational understanding and communication that the chatbot is not an emergency response mechanism — this must be reflected in all public communications about the tool.
No. 23. Unclear duty of care once a CHRD has engaged with the chatbot	
By making a support tool specifically designed for CHRDs publicly available on its website, Child Rights Connect implicitly signals to children that it is a source of help and guidance. This creates a de facto duty of care that may not be matched by the organisation's actual capacity to respond, monitor, or follow up. If a CHRD acts on information provided by the chatbot and is harmed, or discloses a crisis that goes unaddressed, the question of Child Rights Connect's legal and moral responsibility is not straightforward. This risk is compounded by the chatbot operating across multiple jurisdictions with different legal frameworks around duty of care, child protection, and organisational liability, and by the	Prevention measures • Clear plain language terms of use displayed before the child begins interacting — what the chatbot is, what it is not, what Child Rights Connect will and will not do, and what the child can expect. • Honest communication about organisational response capacity — including accurate information about email response times and the fact that the chatbot is not monitored in real time. • The chatbot is covered as an activity under Child Rights Connect's Child Safeguarding Policy, requiring a risk assessment and the application

absence of any terms of use or age verification mechanism that would help define the boundaries of the relationship between the CHRD and the organisation.	of the established safeguarding procedure in the event of any concern or incident reported. Mitigation measures • Designated responsible person for chatbot safeguarding — a named, accountable individual with the authority, resources, and protocols needed to act on safeguarding concerns identified through monitoring or other means, who is either the Child Safeguarding Officer or reports directly to them. • Regular safeguarding review covers the operation of the chatbot — it is not treated as a separate or standalone tool but as part of Child Rights Connect’s broader safeguarding responsibilities. • Incident recording and learning for all crisis disclosures identified during monitoring — enabling the organisation to improve its responses over time.
No. 24. The chatbot provider’s data practices are conflicting with CHRD confidentiality needs	
Child Rights Connect does not own or control the infrastructure on which the chatbot runs. The chatbot provider, as a commercial platform, operates under its own terms of service, privacy policy, and data retention practices, which was never required to be compatible with the confidentiality standards required for a tool serving CHRDs. Specific concerns include: conversation logs being retained by the chatbot provider beyond what is necessary or safe; data being processed through third-party AI model providers such as OpenAI under their own separate data terms; conversations being potentially subject to US legal process including subpoenas or national security requests; and chatbot provider’s data practices changing without notice as the company evolves its business model. Child Rights Connect may have limited or no contractual leverage to require the chatbot provider to meet CHRD-specific confidentiality standards, and CHRDs using the chatbot will almost certainly be unaware of these data flows and their implications.	Prevention measures • Prior to launch, Child Rights Connect reviewed Chatbase documentation regarding data storage and processing. Chatbase processes limited technical data in order to provide the service. Child Rights Connect does not have access to users’ IP addresses within the Chatbase dashboard. Exposure is limited through data minimisation measures including absence of user accounts, restricted dashboard access, and manual deletion of conversations after 3 weeks. • The chatbot uses a closed knowledge base consisting of Child Rights Connect resources and does not browse the open internet. Child Rights Connect does not seek to identify individual users of the chatbot. • Prohibition on use of conversation data for model training. Conversations between CHRDs and the chatbot must never become training data for commercial AI systems. Mitigation measures • Transparent disclosure to CHRDs about data flows and clear explanation about what happens to conversation data.

	• Regular monitoring of Chatbase policy changes. • In case of sudden platform changes, e.g. if a change introduces unacceptable safeguarding risks, the Safeguarding Procedure will be triggered.
No. 25. Model or platform changes by the chatbot provider affecting safeguards without notice	
Child Rights Connect is dependent on chatbot provider as a third-party platform, meaning that changes to the underlying AI model, content moderation policies, safety features, or platform terms of service are outside the organisation's control and may happen without adequate notice or consultation. A model update could alter the chatbot's responses in ways that introduce new inaccuracies, remove existing safeguards, or change the tone and framing of sensitive topics. A policy change could affect data retention, language support, or the availability of certain features that Child Rights Connect relies on for safety. A business decision such as acquisition, restructuring, or discontinuation of the platform could disrupt the chatbot entirely. For a tool serving CHRDs, any of these changes could have immediate safety implications, and Child Rights Connect's ability to detect, respond to, or prevent them is severely limited by its position as a customer rather than an operator of the underlying infrastructure.	Prevention measures • Ongoing monitoring responsibility with safeguarding knowledge is established. Platform changes that appear technical on the surface can have direct safeguarding implications that require a policy response, not only a technical one. • Proactive engagement with the chatbot provider — Child Rights Connect formally communicates in writing to the chatbot provider that its chatbot serves a vulnerable child population with specific safeguarding requirements, and request advance notification of any changes that could affect data practices, safety features, language support, or the underlying model. • Child Rights Connect signs the Data Processing Agreement of the chatbot provider as part of the registration process. Mitigation measures • Post-update testing protocol — whenever a the chatbot provider or model update is identified, a documented testing protocol is activated before the chatbot is considered safe for continued use. This protocol covers: testing responses to crisis scenarios; checking that safety prompts and disclaimers are functioning correctly; verifying that the knowledge base is being drawn on appropriately; and reviewing responses in key languages for quality and accuracy. • In case of sudden platform changes, e.g. if a change introduces unacceptable safeguarding risks, the Safeguarding Procedure will be triggered. • Annual platform review — a formal annual assessment of whether the selected chatbot provider remains the most appropriate platform for the chatbot, covering: whether the chatbot provider data practices continue to meet the standards established in the data processing

	agreement; whether alternative platforms have emerged that offer better safeguarding credentials; and whether the chatbot's growing usage and complexity warrants investment in more robust infrastructure. This review should be reported to organisational leadership. • Six-monthly knowledge base and safeguarding review includes a specific assessment of the platform — reviewing any changes that have occurred since the last review and updating the risk assessment if new risks have emerged.
--	--