



May 2026

How to Design an Ethical AI Chatbot: Key Lessons from Real-World Implementation



Our fifth session in the Equity & AI Learning Series focused on what it actually takes to design ethical AI chatbots in real-world contexts. We were joined by Karina Rios Michel, Chief Creative & Technology Officer at [Girl Effect](#), a nonprofit global organization that designs media and digital tools—including AI chatbots—to support adolescent girls across Africa and South Asia, reaching millions of users. Their platforms now handle tens of thousands of messages daily, creating a significant volume of real interactions—and with it, a depth of learning about what works, what doesn't, and what it takes to build AI systems that are actually safe and useful in practice. Karina leads Girl Effect's work at the intersection of technology, creativity, and social impact, with a strong focus on building tools grounded in the realities of the communities they serve. The session was highly practical, with lots of thoughtful questions and discussion throughout. What stood out most was the level of focus, time, and intentionality behind the work. This is not quick or easy work. It reflects a long-term investment in collaborative design, continuous learning, and iteration. That depth of effort is what has enabled real scale and impact, and what made the lessons shared in this session so valuable for other organizations thinking about how to use AI and chatbots in meaningful, responsible ways.

Why This Session Matters

Very few organizations are investing the time, funding, and long-term effort required to actually build and test AI tools in real-world contexts—especially in ways that are safe, ethical, and grounded in community realities. What made this session particularly valuable is that [Girl Effect](#) has done exactly that, and has chosen to share the journey openly. Throughout the discussion, it was clear that this work has taken years of investment: building systems, testing models, learning from failures, and continuously adapting. It also highlighted a broader challenge in the sector—while many nonprofit organizations are exploring AI, far fewer are investing in research, documentation, and sharing lessons in a meaningful way.

Karina spoke directly to this tension. There is growing interest in AI, but limited shared learning, fragmented research, and increasing pressure around competition and protecting intellectual property. In that context, Girl Effect is taking a different approach by prioritizing openness, collaboration, and contributing to a broader evidence base on what works and what doesn't. That shift matters. It creates space for more grounded, responsible use of AI across the sector, especially for organizations working in complex, resource-constrained, and equity-focused contexts.

What We Explored Together

AI IS NOT NEUTRAL ESPECIALLY FOR GIRLS.

AI is advancing rapidly, but it risks being built in ways that leaves entire cultures, languages, and the voices of girls and young people behind.



Karina walked through the full journey of building Girl Effect’s chatbots, from early experiments to large-scale deployment, and what it actually takes to make these tools useful, safe, and relevant in practice. A few things came through clearly:

This work starts long before AI

The chatbot didn’t start as a chatbot. It evolved over more than a decade—from Girl Effect’s magazine-style advice platforms (like Ni Nyampinga Magazine’s “Baza Shangazi” (“Ask Aunty”) advice column in Rwanda) to SMS and radio, to digital platforms, and eventually to AI-powered chatbots. This is important because the foundation wasn’t technology—it was listening. Years of questions from girls and young people. Listening about relationships, health, identity, and daily challenges. This all shaped what the chatbot eventually became. The key point here is that if you don’t deeply understand the questions people are actually asking, AI won’t solve that gap—it will amplify it.

AI is one layer within a broader ecosystem

Girl Effect’s chatbots (such as Big Sis and Bol Behen bots) are not standalone tools. They are part of a wider ecosystem that includes media brands, social content, radio, and community engagement. Young people typically encounter the chatbot through these existing channels, in content they already trust, follow, and engage with. The chatbot then becomes a continuation of that relationship, not a new or unfamiliar entry point. This matters because trust is not built through the chatbot alone. It is built over time through consistent, relevant, and culturally grounded content. The chatbot works because it sits inside that trust—not outside of it. Karina emphasized that without this ecosystem:

- users may not engage at all
- sensitive questions may not be asked
- or responses may not be trusted

This means that launching a chatbot without an existing engagement channel (audience, platform, or trusted brand) could significantly limit its usefulness (particularly for youth). The question is not just “how do we build a chatbot?” but “where does this fit within how people already access information and support?” It also shifts how we think about impact: the value is not just in the chatbot interaction itself, but in what happens because of it—whether someone learns something new, makes a decision, or is able to access support or services they wouldn’t have otherwise. In Girl Effect’s case, the chatbot is often one step in a longer journey. A conversation might start with a question, but it can lead to deeper understanding, behaviour change, or being connected to real-world support. The impact is not the interaction—it’s what the interaction enables.

Human support is built in from the beginning

A core design decision was keeping humans in the loop from the beginning. Karina drew a clear line between what AI can do well and where it cannot go further. AI is effective at answering common questions, providing information at scale, and helping users understand their options. But when situations become more personal, emotional, or high-risk, information alone is not enough.



This is where human support becomes essential. Girl Effect intentionally designed their system so that users can be escalated to trained agents who can provide reassurance, context, and guidance in ways that AI cannot. Around 30% of interactions are escalated to trained human moderators or counselors, especially in all cases involving abuse, violence, or sensitive disclosures. A significant part of the work was defining where AI is appropriate, and where it is not. This includes: not generating certain types of advice, not responding to situations involving risk or harm and not speaking on behalf of lived experience.

Co-design is continuous and deeply contextual

These chatbots are co-designed with young people and adapted to each context, including language, tone, cultural norms, and even humor. Karina shared examples of training models to understand local language patterns like Sheng (in Kenya) or mixed-language inputs, and ensuring responses feel natural and not scripted or external. Without this level of adaptation, AI tools may function technically, but they will fail to resonate, build trust, or be used at all.

Scale is built through systems, not just technology

Girl Effect's ability to respond to tens of thousands of questions each day is not just about using AI, it reflects years of building systems around it. This includes moderation and safeguarding processes, escalation pathways to human support, partnerships with service providers, and continuous testing and refinement of content and responses. These systems are what make it possible to use AI safely at scale, especially in contexts involving sensitive topics like health, relationships, and violence. Karina emphasized that this infrastructure takes significant time, investment, and coordination. It is not something that can be quickly replicated by simply adopting a tool or model. Designing the technology is only one part of the work. The larger work, and the real responsibility, sits in designing the systems, partnerships, and safeguards that make AI safe and meaningful in practice.

Key Lessons: What Works in Practice

A number of clear, practical lessons emerged from Karina's experience which are particularly relevant for organizations thinking about whether and how to build AI tools in their own contexts.

- 1. Start with a real, high-volume need, not the technology:** Girl Effect didn't start with "let's build a chatbot." They were already receiving large volumes of questions from young people—many of them deeply personal and difficult to answer at scale. The chatbot emerged as a response to that pressure: how do we respond to thousands of real questions in a way that is timely, consistent, and safe? *The takeaway: AI is most useful when it responds to a clear, existing demand—not when it is introduced as a new solution looking for a problem.*
- 2. Iteration is constant, and necessary:** The chatbot is not a fixed product. It is continuously tested, reviewed, and updated based on user interactions, flagged responses and changing contexts and needs. Karina shared that this requires ongoing investment - teams reviewing outputs, refining responses, and improving the system over time. *The takeaway: this is not a one-time build. It is an ongoing process that requires high level and continuous resourcing.*
- 3. Trust is built through consistency, not just accuracy:** Accuracy matters—but what matters just as much is whether responses feel relevant, respectful, and aligned with the user's reality. This comes from tone and language, cultural relevance and consistency across interactions. *The takeaway: trust is not just about getting the answer right—it's about how the answer is experienced.*

4. **Define success based on outcomes, not chatbot usage:** One insight shared was the distinction between engagement and impact. A chatbot can generate high volumes of interaction—thousands of messages, repeat users, long conversations—but that does not mean it is achieving meaningful outcomes. Girl Effect is very intentional about this. The goal is not to maximize conversations, but to support progress. That might mean helping someone better understand a situation, feel more confident in a decision, or take a step toward accessing support. In many cases, the real impact happens after the interaction. *The takeaway: Success is not how much the tool is used, but what changes because of it. This has significant implications for user metrics and how to assess impact.*
5. **Design for how people actually want to engage:** Karina also shared that engagement doesn't come from simply offering a chatbot, it comes from designing an experience that reflects how people actually interact. Rather than relying on a fully open, conversational model, Girl Effect designed a hybrid experience. Users can ask questions directly, but they can also navigate structured content through menus. This gives people flexibility; some want to explore, others want to ask something specific. They also paid close attention to tone and usability. The chatbot is designed to feel relatable and peer-like, not clinical or overly technical. This includes using familiar language, culturally relevant references, and a simple, intuitive interface. *The takeaway: engagement is designed. It comes from understanding user preferences and creating an experience that feels natural, flexible, and easy to use.*
6. **Be explicit about what your AI can and cannot do:** Another important lesson is how much effort goes into setting expectations with users from the start. Girl Effect makes it very clear that the user is interacting with a chatbot, what it can and cannot do, and where its limitations are. This is not just a technical detail—it is a core part of ethical design. Being transparent helps prevent over-reliance, reduces the risk of misunderstanding, and ensures that users are not placing inappropriate trust in the system. *The Takeaway: Clarity and transparency are essential to safe and responsible use.*
7. **Design intentionally to avoid user dependency:** One of the most distinctive aspects of Girl Effect's approach is that the chatbot is not designed to become a long-term emotional support system. Instead, it is positioned as a bridge—helping users understand their situation and guiding them toward real-world support. This is reflected in how the chatbot communicates, how conversations are structured, and how users are encouraged to seek additional help when needed. In this sense, Girl Effect is not optimizing for engagement in the way many digital AI platforms do. The chatbot is not designed to keep users in conversation, but to help them move forward. A successful interaction might mean that a user leaves the chatbot—because they feel informed enough to take action, speak to someone, or access a service. This reflects a very different understanding of success: not retention, but progression. *The takeaway: Design your tool to support independence, not reliance.*
8. **Build safety through layered safeguards, not a single feature:** Karina emphasized that safe AI does not come from one feature or safeguard, but from multiple layers working together. This includes using vetted and expert-approved content, setting limits on what the AI can generate, monitoring interactions for risk, and having clear escalation pathways to human support. These layers are continuously tested and refined, based on real interactions and emerging risks. *The takeaway: Safety must be designed across the entire system, not added at the end.*
9. **Design your chatbot to model respectful behavior, not just respond to it:** One of the more unexpected lessons Karina shared was how users sometimes engage with the chatbot using abusive or inappropriate language. Rather than simply filtering or ignoring this, Girl Effect made a deliberate design choice: the chatbot responds in ways that model respectful behavior and reinforce boundaries.

Instead of complying or continuing the conversation, the chatbot can push back—making it clear when language is not appropriate, and reinforcing norms around respect, consent, and communication. In doing so, the chatbot becomes more than just a source of information, it becomes part of the learning experience. This reflects a broader shift in design: not just optimizing for responsiveness, but intentionally shaping how interactions happen. The chatbot is designed to reflect the values Girl Effect is trying to promote—confidence, self-respect, and healthy boundaries. *The takeaway: AI tools are not neutral in how they respond. They can reinforce harmful dynamics—or actively model the behaviors and values you want to see.*

10. Plan for long-term investment, not quick wins: One of the most important realities shared in the session is the level of investment behind this work. What looks like a chatbot is actually the result of more than a decade of program development and years of dedicated AI work. This includes building datasets grounded in real user questions, developing and testing content, creating partnerships with service providers, governments and communities, and continuously refining the system. *The takeaway: Effective, ethical AI systems require sustained effort, iteration, and learning over time.*

11. Invest in real, local data. Don't rely on generic AI language models: A critical part of Girl Effect's approach is the depth of investment in language, culture, and context. Rather than relying on generic AI systems and translating outputs, they built their own AI layer trained on millions of anonymized messages from girls across the countries where they work. This allows the chatbot to reflect how young people actually speak, including slang, mixed languages, tone, and cultural references. In practice, this means the chatbot doesn't just translate language, it understands context, nuance, and meaning in a way that feels relevant and trustworthy. This level of relevance does not come from technology alone. It comes from years of collecting real questions from users, building relationships with communities, working with local advisors and experts and continuously adapting content to reflect lived realities. Karina emphasized that without this investment, AI tools risk being technically functional but culturally irrelevant or even misleading. *The takeaway: if your AI is not grounded in real, local data and lived experience, it will not resonate, and may do more harm than good.*

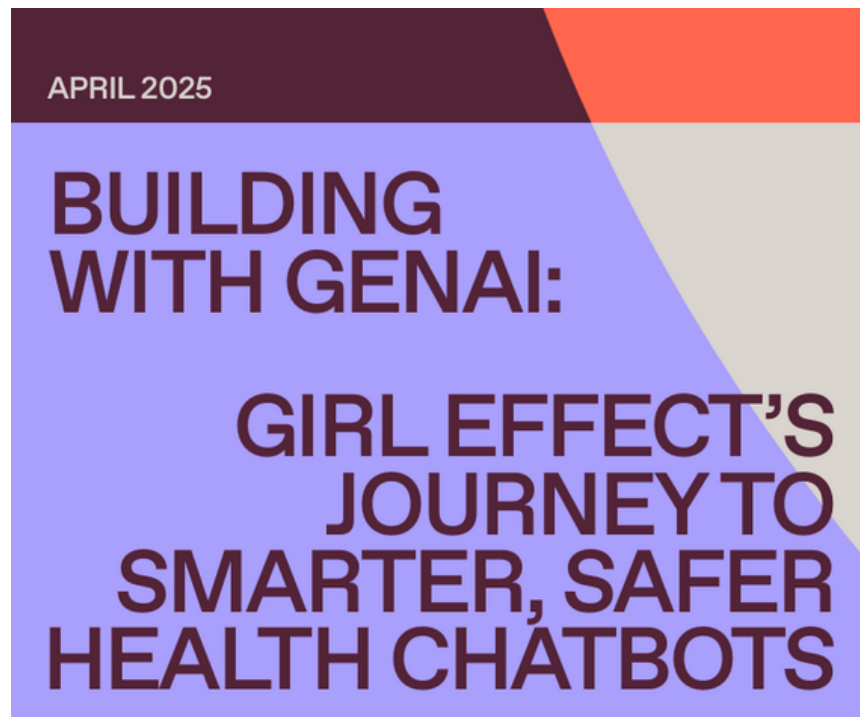
Resources

Below are a few key resources for those interested in exploring this work further:

- Girl Effect: Overview of Girl Effect's work designing media and digital tools—including AI chatbots—focused on supporting adolescent girls across Africa and South Asia. <https://www.girleffect.org/>
- AI Chatbot Demo (Multilingual Sandbox): A live demo environment shared during the session that allows you to explore how different models and prompts perform across languages and contexts. This reflects the type of testing environment used by Girl Effect teams to iterate and refine chatbot design. <https://genai.girleffect.org/chatbot/>

**GIRL EFFECT
ETHICAL AI
GUIDELINES**
**FOR THE DEPLOYMENT OF
SOCIAL AND BEHAVIOUR
CHANGE CHATBOTS**

- Girl Effect AI & ML Whitepaper: A deeper dive into their approach, including system design, implementation decisions, and lessons learned from deploying AI chatbots at scale in real-world contexts.
 - Building with GenAI: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ge-prd-eu-west-1-webapp-assets-01.s3.eu-west-1.amazonaws.com/girleffect/media/documents/Girl-Effect_AI_ML-Whitepaper.pdf
 - AI and ML Vision: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://girleffect.org/static/theme/Girl_Effect_AI_MLvision_2024v2.pdf



Closing Reflection

This was a really inspiring session, both in terms of the depth of experience shared and the openness around what it actually takes to design ethical AI in practice. There was strong engagement throughout, with thoughtful questions and reflections that highlighted how relevant—and complex—this work is across different contexts. What stood out most is that this work is possible. But it requires intention, investment, and a willingness to learn and adapt over time. Karina and the Girl Effect team offered a rare and valuable look into what that really looks like in practice.

If you have ideas for future session topics, potential speakers, relevant case examples, or questions you'd like us to explore, I would love to hear from you. Please feel free to reach out to me - Paula Richardson via email at richardson@salanga.org or connect with me on linkedin - <https://www.linkedin.com/in/paula-nicole-richardson/>.