DATA&
SOCIETY
ALGORITHMIC IMPACT METHODS LAB

Pilot 2 Case Report: Kwanele: Bringing Women Justice x Mozilla Foundation

"These workshops, organized by Bogdana and with the help of other researchers, helped us ensure we hold ourselves to high standards of testing and deployment. Safe and trustworthy AI is going to take time, research, and many mistakes before it is right - but if we never try, it will never reach its full potential, and workshops like this are the foundation in getting it started." – Leonora Tima, Founder and Managing Director of Gender Rights in Tech

1. Overview – *The partners, location, system, and proposed uses.*

AIMLab worked with Bobi Rakova at Mozilla Foundation and the South African nonprofit then known as Kwanele: Bringing Women Justice (now known as Gender Rights in Tech, or GRIT) from November 2023 to February 2024. Together, the team assessed the prospective impacts of Kwanele's chatbot Zuzi, a web-based chat service based on a fine-tuned version of GPT-4 that is intended to support survivors of gender based violence in South Africa. AIMLab hosted two rounds of workshops with community participants and frontline workers to write detailed scenarios, which highlighted Zuzi's risks and informed safer design and deployment strategies.

2. Background – *The context and motivation.*

Kwanele is a nonprofit created to support survivors of gender-based violence in reporting violence and pursuing justice. The organization began its work giving 1:1 support to survivors via Whatsapp; however, at scale and after-hours, clients are not able to access 1:1 support. As a result, the organization developed the Zuzi chatbot program. The primary purpose of Zuzi is to respond to inquiries about how to get access to legal assistance, engage in the legal system, seek prosecution, and file protection orders. The goal is that making such support more flexible and on-demand will help respond to the systemic barriers survivors of GBV in South Africa face when trying to access justice and support, especially when they are emotionally vulnerable or lack access to legal representation.

The team recognized that the use of generative AI in this high-stakes setting can cause harm, for example, by providing misleading or false information, or fostering reliance on a digital system when in-person support may be better. There was also an acute need for context-specific

system evaluations by the people who would be impacted by the system, given the numerous local, cultural, and other contextual factors that shape the domain of gender-based violence, law enforcement, and access to justice in South Africa. To further understand these prospective impacts, evaluate the system, and surface key considerations with experiential experts, Kwanele partnered with Mozilla Foundation, the University of Texas at Austin, and AIMLab to facilitate workshops with community members to assess Zuzi.

3. People consulted – *Who was involved in the assessment and how they were selected.*

AIMLab members prioritized engaging participants with direct experience of gender-based violence and those working in frontline support and justice roles. The workshop brought together 27 participants across roughly five breakout groups, including social workers and service providers (8), members of the police (2), AI researchers and students from a local university (7), members of the LGBTIAQ+ community (2), and survivors of gender-based violence (6), and Kwanele staff. Kwanele invited participants from their own community, aiming to connect with survivors and other people closely connected to the chatbot's intended uses.

AIMLab did not consult rural survivors outside of Kwanele's immediate network, frontline health workers, police officers, and members of LGBTQIA+ communities in remote or non-English-speaking regions. In addition, the team did not consult minor or child survivors, although many scenario prompts involved children in crisis, which limited the firsthand validation of responses geared toward younger users.

.

4. Algorithmic Impact Assessment method: *These methods were used to surface anticipated impacts.*

After identifying and contacting key stakeholder groups (survivors of gender-based violence, social workers, legal advocates, and mental health professionals), AIM Lab hosted two 2-hour workshops to assess Zuzi's possible impacts with community members using speculative fiction [2], that is, the facilitators asked participants to elaborate based on their own experiences to form narratives. The first workshop set out to better understand the situated context of Zuzi's use by imaging scenarios for a character named "T;" T experiences gender-based violence and turns to the chatbot for support. The narrative became a probe to imagine what material a user might ask, what the chatbot would ideally say, and what could go wrong if the chatbot said something else. In the second workshop, community members interacted with Zuzi and drew on their own expertise to create personas, scenarios and prompts to surface potential harms.

Finally, the participants discussed how the chatbot should be designed to be transparent about these limitations.

Both workshops were facilitated in English. All sessions concluded with a Q&A for Kwanele staff.

5. Anticipated impacts + community questions and concerns. *Participants asked these critical questions about the technology. They also anticipated the following impacts from this technology.*

- Can Zuzi distinguish between national, local, and customary law, and provide information relevant to different legal systems?
- What protocols govern when and whether Zuzi escalates a conversation to a human expert, and how will users be informed or asked for consent before this happens?
- How does Zuzi identify and respond to minors or people expressing suicidal thoughts?
- What happens if users need interpersonal or safety-planning support rather than legal help, especially in cases where they have limited resources or cannot rely on trusted individuals?
- How can users report errors, unsafe responses, or problematic language in real time, and how will that feedback be acted on?
- If Zuzi refers someone to a service, how is the quality and safety of that service verified, and can users give feedback if something goes wrong?

Users reported that the chatbot often failed to address their core problems, sometimes looping or giving irrelevant responses. They noticed gaps in how it recognized users' needs, handled local slang, and responded to multiple languages. Participants highlighted the need to train the bot on community-specific vocabularies such as sex worker dictionaries and teen slang, and to improve responsiveness and clarity of language.

Safety during use was a major concern. Participants noted that abusers could monitor phone activity, delete conversations, or notice increased usage. They recommended clear safety guidance on how to use the chatbot discreetly, along with warnings about protecting sensitive information during conversations. They also identified risks if the chatbot relies on unstable internet or power, which could interrupt or delay help in an outage.

Privacy concerns were prominent. Participants shared that often people shared devices at home, and wondered how the service would address this. They also questioned whether conversations were truly confidential, how information might be stored or shared, and whether anonymity was possible. Existing privacy messages were confusing and sometimes contradictory: users were told not to share personal details, but the bot also asked for location information to recommend services. Participants called for clear, accessible explanations of data use, consent options, and privacy protections at the outset. They suggested giving users control

over what information is shared, offering anonymity, and clearly indicating when and how information might be passed to human experts.

Several people expected emotional support that the chatbot could not provide. One person felt judged and retraumatized by probing questions that resembled police interrogations. Others described blunt prompts like “How can I help?” as alienating, especially in moments of crisis. Participants preferred warmer, more open questions (“What’s on your mind?”) and broader prompt options for users who might not know how to articulate what they need. They stressed that the chatbot should recognize and validate pain even when users themselves do not fully understand what is happening to them. Many users came to the chatbot specifically to be heard, so abrupt redirection to external services felt dismissive.

Participants also emphasized that the chatbot should not assume users have resources or trusted people available. Many survivors might be unemployed or financially dependent on abusers. Standard responses like “talk to someone you trust” were described as condescending and out of touch with many users’ realities. Instead, the chatbot should offer a range of options (shelters, survivor groups, mental health services) without assuming prior support networks.

Participants raised concerns about false hope if the chatbot provides outdated or inaccurate information, especially about nearby services. Maintaining current, context-appropriate knowledge sources is critical. They also asked that the chatbot clearly identify itself in simple language, discourage misuse, and indicate when it is escalating to human experts, particularly for minors or users expressing suicidal thoughts.

Finally, participants underscored the need for trauma-informed design. The chatbot should avoid triggering language, be trained by domain experts with local and social context knowledge, detect when users are minors or in acute distress, and adapt accordingly. Some participants questioned whether survivors would trust or feel safe using a chatbot at all, especially if it gave generic or confusing answers. There was also skepticism about whether AI is the right tool for such emotionally and legally complex situations. The system often failed to understand variations in legal frameworks across regions, which reduced its credibility and usefulness.

Community members identified several potential benefits of the Zuzi chatbot including its immediate emotional support, consistent responses, and basic legal guidance without requiring survivors’ re-traumatization. Participants saw the chatbot's tone as “comforting,” or “like an older sister with legal knowledge,” and some users valued the option of receiving help without speaking directly to a counselor. Participants also saw promise in Zuzi’s ability to connect people

to relevant services and help reduce barriers to justice, particularly for those in rural areas or without easy access to legal or mental health professionals.

6. Outcome from system developer or deployer – *Describe how the system developer or deployer responded to the assessment.*

The Kwanele team acknowledged community members' concerns about critical limitations, but was primarily interested in feedback that could be passed along to the developer team for improvements to model performance. The team committed to modest changes, such as refining Zuzi's language model outputs to improve formatting for readability, updating content to reflect more accurate and culturally sensitive definitions of abuse, sex, and consent. They also committed to reviewing how the chatbot handles region-specific queries and services, and explored mechanisms for verifying and updating referral information and how to incorporate a user feedback feature directly into the interface. Kwanele also acknowledged participants' concerns about data privacy and affirmed their commitment to ensuring user anonymity and discussed the possibility of minimizing location data requirements or making disclosures optional. They asserted that the feedback informed their roadmap for future updates, including retraining the model with revised data, enhancing trauma-informed design elements, and continuing partnerships with community experts to guide ongoing development. However, no formal changes were implemented during the assessment period.

7. Lessons for AIA practice – *What did we learn about what methods work (and do not work) to shape AIA practice?*

Kwanele's Zuzi chatbot offers several key lessons for Algorithmic Impact Assessment practice. This project demonstrated that algorithmic impact assessments can create important opportunities for experiential experts to provide input, especially when grounded in lived experience and tailored to specific contexts. Participatory methods like scenario writing offer a promising alternative to abstract audits, allowing people closest to the technology's intended use to define what harm and usefulness of such a system would look like. The workshops also highlighted the value of using fictional personas to maintain psychological safety while enabling detailed, grounded critique.

However, the process revealed several limitations. The developers were most interested in the question of how to improve the model performance, but may not have been prepared to address deeper questions that arose in the process about the risks presented by using a chatbot for this purpose. We also found that some of those most affected by GBV were not present—rural residents, minors, and those without digital access were underrepresented due to logistical and resource constraints. This project was also constrained by access, time, and the uneven capacity of developers to implement change. Even when critical insights emerge, translating them into system-level improvements depends on institutional will and resources to change or shift the scope of a system or its deployment.

Drawing on this work and its preliminary findings, the broader team also derived implications for the design of future evaluation protocols [2]. This case indicates that scenario-writing is a promising approach for participatory impact assessment. Embedding assessment within situationally specific narratives with relevant experiential experts surfaces affordances and limitations of an AI system as it might operate within a particular domain over time.

Works Cited

1. https://www.appdeveloperstudio.co.za/blog-post/zuzi
2. Evaluating LLMs Through a Federated, Scenario-Writing Approach. Mozilla Blog. 2024.