# SAMESAME

Patrick J McGovern FOUNDATION

02/28/24

# FINAL INSIGHT REPORT

# OVERVIEW

# LEVERAGING AI TO SUPPORT LGBTQI+ MENTAL HEALTH

Thanks to the support of the Patrick J McGovern Foundation and our inclusion in the Data & Society Accelerator Programme, the team at SameSame have significantly upgraded our technical infrastructure and dramatically changed the way we collect and use data to design and deliver impactful digital mental health services to LGBTQI+ youth in South Africa and Zimbabwe. The Reinforcement Learning (RL) system we've built, the main product of our work with PJMF, is not without its challenges and it will continue to undergo changes now that the Accelerator Programme has ended, but the data we've collected has helped us improve the number of young people we're able to reach and the percentage of those young people who go on to have meaningful experiences engaging with our chatbot.

# THE NUMBERS

85k+

More than 85k users have started conversations with our chatbot in Zimbabwe and South Africa over the course of the Accelerator programme

71%

71% of users who complete the Patient Health Questionnaire (PHQ8)—a standardised clinical tool delivered through our chatbot—test positive for depression

63%

63% of users who complete the baseline and endline surveys (PHQ8) after engaging with our content show improvements in their mental health

8

Through the introduction of the Reinforcement Learning system we've built through the accelerator, we've launched 8 different experiments that are continuously running.

2x

Through these experiments we've doubled the number of users who successfully complete sessions of our CBT course.

SAMESAME

Patrick J McGovern FOUNDATION

# THE CONTEXT

SAMESAME

Across the globe, LGBTQI+ youth are disproportionately affected by mental health challenges. In South Africa, for example, LGBT youth experience depression at 4x the rates of their peers, despite legal protections.

Even when mental health services are available, stigma and prejudice limit access — more than 50% of health workers in South Africa consider gays and lesbians 'disgusting', limiting access to essential health services.

/ The South African Depression and Anxiety Group, SADAG, 2016

SAMESAME

The last few years have seen a rapid increase in the range and availability of digital mental health support services, but virtually **none of them cater to the specific needs of LGBTQI+ youth** in low- and middle-income countries, where device compatibility, data costs and data privacy concerns blocks access and where a lack of contextually-relevant, culturally-responsive content inhibits their engagement.

At SameSame, we know that **context matters.**

Our mission

## SameSame was founded to develop and scale safe, confidential and identity-affirming mental health support to LGBTQI+ youth across the world, especially in places where it's difficult or even illegal to be different

# Our Approach

I

Identify interventions that have been proven to boost the mental wellbeing of LGBTQI+ youth — and adapt them to local contexts and for digital engagement.

II

Test novel approaches to the dissemination of these products through safe and secure digital channels, like WhatsApp, in places where offline support is scarce or daunting to access.

III

Create service referral pathways to partners, like grassroots community organizations, expanding their capacity to reach and serve LGBTQI+ youth and improve their health.

SAMESAME

# EVIDENCE-BACKED FOUNDATIONS

For our first product, we've partnered with the creators of a Cognitive Behavioral Therapy (CBT) intervention called AFFIRM.

**In multiple peer-reviewed trials**, AFFIRM has been demonstrated to **reduce depression by 63%** and improve knowledge and awareness of coping skills by 57%.

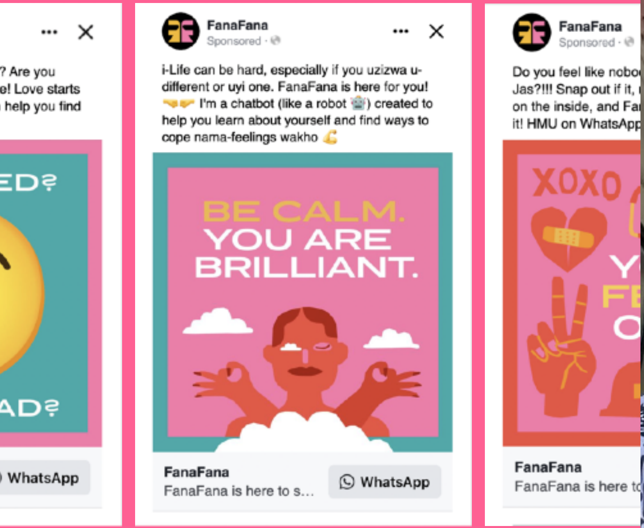
†www.projectyouthaffirm.org

# SCALABLE DIGITAL DISSEMINATION

We've adapted AFFIRM so it can be delivered via a WhatsApp chatbot. Whatsapp is the perfect delivery channel because:

🌍 Ubiquitous and easily accessible
🤗 Our users love it and trust it
💰 Light on data, so it's affordable
🔒 Encrypted, so it's safer

# We reach youth through...




**MEDIA**

Facebook, Instagram, Google Ads, Tinder, Grindr

**AMBASSADORS**

Volunteer and paid online ambassadors

**PARTNERS**

Outreach through on-the-ground LGBTQI+ orgs

# THE USER JOURNEY


LGBTQI+ youth access on-the-ground services (like HIV testing) and, depending on their needs, are referred to our chatbot as part of a ‘stepped care’ approach.

LGBTQI+ youth interact with the bot and complete the AFFIRM course, equipping them with the tools to fight depression and anxiety and improve their mental wellbeing.

Chatbot users receive regular, affirming messages from the bot, helping them maintain their mental wellbeing and encouraging them to share the bot with their friends.

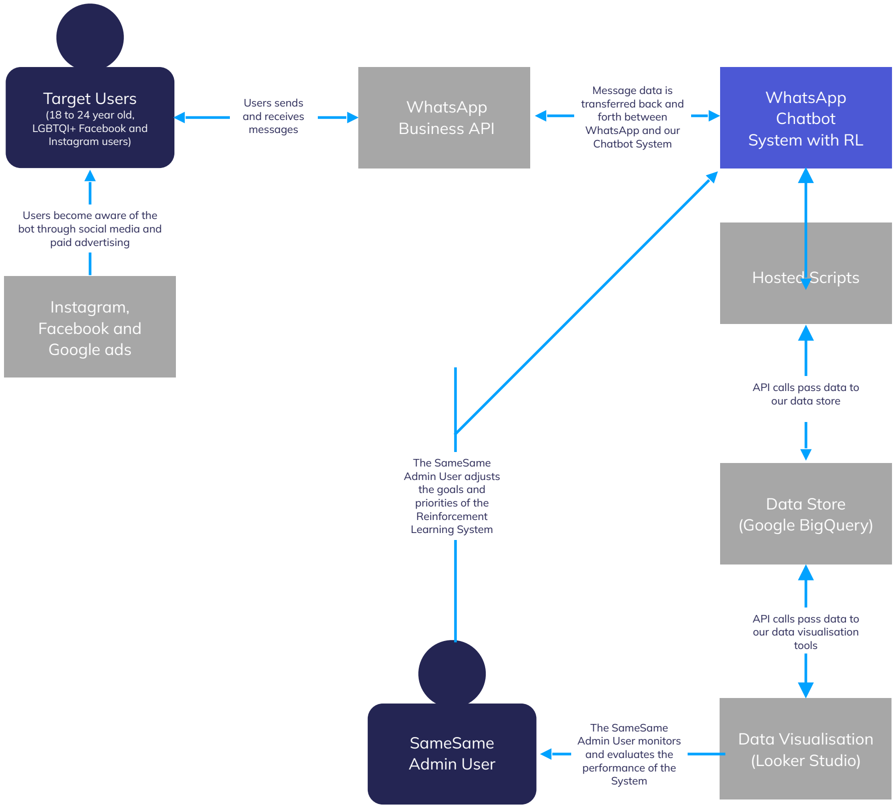
As users interact with the bot, we gather data and, with the support of machine learning tools, assess their needs to promote other services they might access to improve their health, education or employment opportunities.

Based on our assessments or their direct requests, users are connected to offline service providers who track the source of the referral and report back to the SameSame team.

# PROJECT BACKGROUND

SameSame is a small team with big goals and limited bandwidth. We want to make secure, confidential, identity-affirming mental health support available to as many LGTBQI+ young people as possible, as quickly as possible — using our chatbot, delivered through WhatsApp and Facebook Messenger. But there's limited evidence with regard to exactly what will deliver the desired results in the places we work. Limited resources constrain our ability to run, assess and learn from experiments that seek to understand what mix of content and user experience designs will drive higher engagement with our bot and, ultimately, greater improvements in mental health among our users. **Reinforcement Learning (RL)** is concerned with maximizing the desired outcomes in a particular situation using an automated trial-and-error method, rather than drawing on existing data inputs. RL uses algorithms that learn from outcomes produced through multiple, concurrent 'trials' and decide which action to take next. Our hypothesis for the project we launched through the Accelerator is that we can use RL to become much more efficient at continuously adapting our chatbot to evolving local conditions, removing the need for us to make many small, time-consuming decisions ourselves.

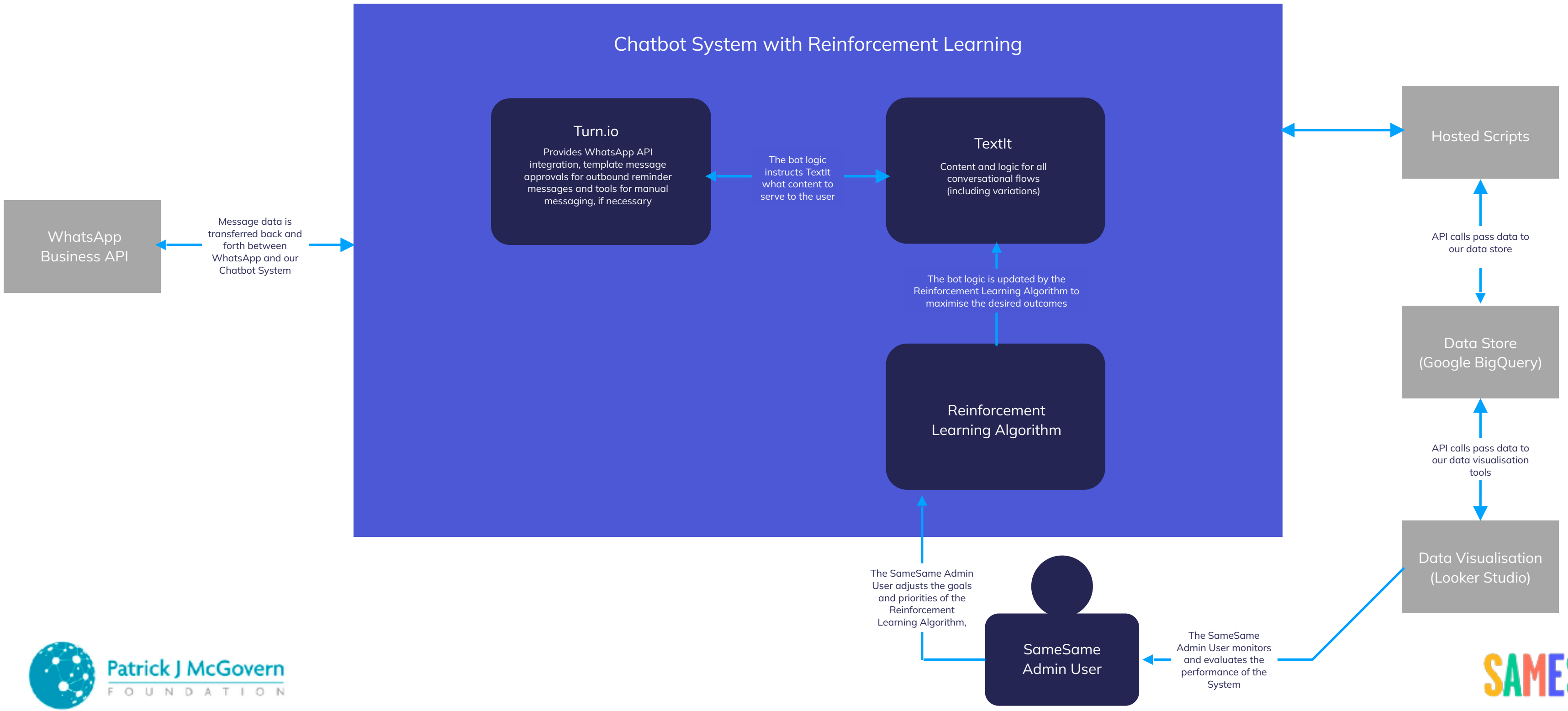
# PROHECT SYSTEM CONTEXT




Patrick J McGovern FOUNDATION

SAMESAME

# PROJECT SYSTEM COMPONENTS




Patrick J McGovern FOUNDATION

SAMESAME

# OUTPUTS

SAMESAME

# OUTPUTS

Our inclusion in the PJMF Data & Society Accelerator programme allowed us to develop a range of different outputs that support SameSame's more efficient and effective use of data.

**01**

**RL System**
A reinforcement learning system that allows us to continuously run multi-arm experiments across our chatbots

**02**

**A new dashboard**
Working with a data scientist, we were able to build a dashboard that provides the SameSame team with more reliable, readily available information.

**03**

**Data Security Audit**
With the Cyberspace Institute, we were able to conduct a rapid security audit of our infrastructure and systems to identify areas that require more attention.

**04**

**User Research**
Using the data gathered through this project we were able to define a set of questions we asked users in 1:1 interviews to better understand the data we were seeing on our dashboard.

SAMESAME

Patrick J McGovern FOUNDATION

# 01 REINFORCEMENT LEARNING SYSTEM

They key output of our work with PJMF was the development of a **Reinforcement Learning (RL) system** that allows SameSame to run multiple versions of each of our chatbot's 'conversational flows', continuously comparing outcomes across these variations to automatically determine which flows are most likely to deliver impact in any given context.

The SameSame team uses TextIt to build the conversational flows that structure a user's experience of the chatbot and Turn.io to connect to the WhatsApp Business API. Data collected from TextIt and Turn.io is transferred to Google BigQuery and, from there, the data is used to populate dashboards that have been set up on Google Looker Studio.

SAMESAME

# 01 REINFORCEMENT LEARNING SYSTEM

We connected the different components of our system to allow for context-driven user experiences and experiments on the user-facing frontend (Glific) and data analytics and insights on the backend (DDP). The frontend now supports the tagging of user sessions that are involved in an experiment with the properties that include the Experiment ID (referencing the the experiment the user was a part of); Treatment Arm ID (the treatment assigned and delivered to the user (associated with one experiment ID); and the timestamp of the treatment assignment and delivery. This information is combined with user session outcomes and metrics that are defined in the backend (DDP) to determine the performance of the different treatment arms and run statistical tests against these outcomes to determine their statistical significance.

SAMESAME

# 01 REINFORCEMENT LEARNING SYSTEM

We've built a data application on top of the DDP backend that maintains a count of the successes and failures for each outcome related to the treatment arms of each experiment, probabilistically deciding which treatment arm to pull next, based on the performance-to-date of each arm. When the next user session is initiated into a specific experiment, an API call is made to this data application to decide what treatment arm the next user will experience and the loop continues. This approach is known as the Beta-Bernoulli Multi-armed bandit and draws upon Thompson Sampling to guarantee the minimization of "regret" in trading of exploration against exploitation.

A code notebook providing an example of this concept can be found here. Please find more details on the experiments please see the following slides.

SAMESAME

# 01 REINFORCEMENT LEARNING EXPERIMENTS

The first experiment we conducted was an experiment designed to discover what factors were likely to drive completion of our chatbot's onboarding experience — making sure that users got to the point they were ready to start the first session. Wen created a 5-arm experiment.

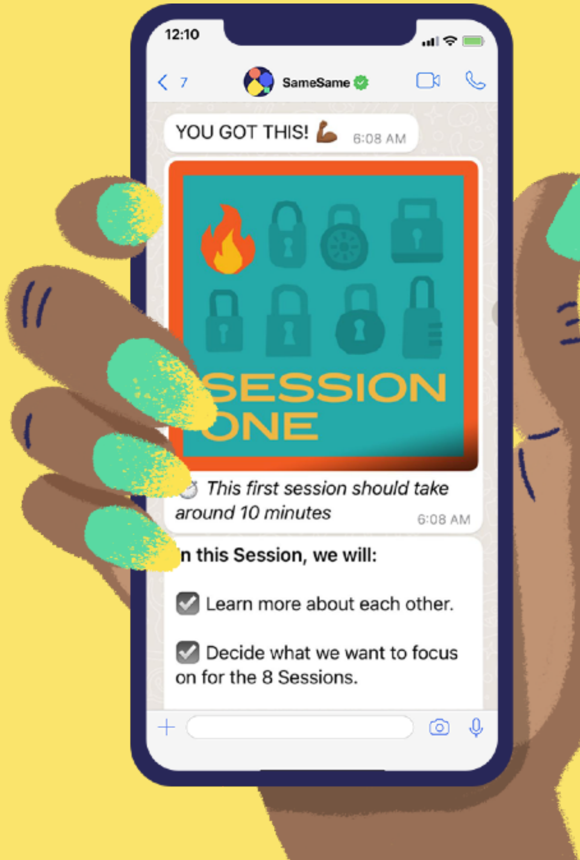
**01**
The original onboarding experience

**02-A**
An onboarding experience which used informal language and colloquialisms - a 'friendlier' version

**02-B**
A version of the friendly onboarding which was much shorter, half the length of the original

**03-A**
An onboarding experience with a focus on the 'medical' benefits of the bot, using a reassuring but authoritative tone

**03-B**
A much shorter version of the onboarding with the authoritative tone



# 01 REINFORCEMENT LEARNING EXPERIMENTS

Based on the most recent data from are ongoing onboarding experiment, arms 02-B and 03-B are the most successful. Completion for 01 stood at 46% compared to between 96% and 86% for 02-B and 03-B respectively. One of the things we've determined through this experiment is that, especially in the earlier stages of the chatbot experience, content is less relevant to users than the brevity and clarity of the messages they receive. This chimes with guidance we've received from one of the co-founders of turn.io to build experiences 'for minimal engagement' — in other words, design experiences that require as little effort from users as possible.

That said, one good reason to keep the experiment running is to track whether or not there are any correlations between the onboarding experience the user has and the likelihood they complete the CBT course. We may discover, over time, that the authoritative arms may limit the number of users who initially engage but that these users are more likely to engage more deeply with the content. This is one advantage to the RL system — we can modify the outcomes each experiment optimises for over time as we learn more about how users engage with the entire experience rather than just individual flows.

**01**
Original

**02-A**
Friendly

**02-B**
Friendly short

**03-A**
Authoritative

**03-B**
Authoritative short

# 01 REINFORCEMENT LEARNING EXPERIMENTS

Other experiments that are currently live (and their results) are highlighted below.

Comparing different approaches to incentivise users to sign up for push notifications

82% (2180) users opted in for push notifications

Comparing different discovery routes for the CBT sessions

Improved Session 1 completion from 6% to 20% for all users that initiative conversations with the bot

Comparing linear with user-choice paths through the CBT course

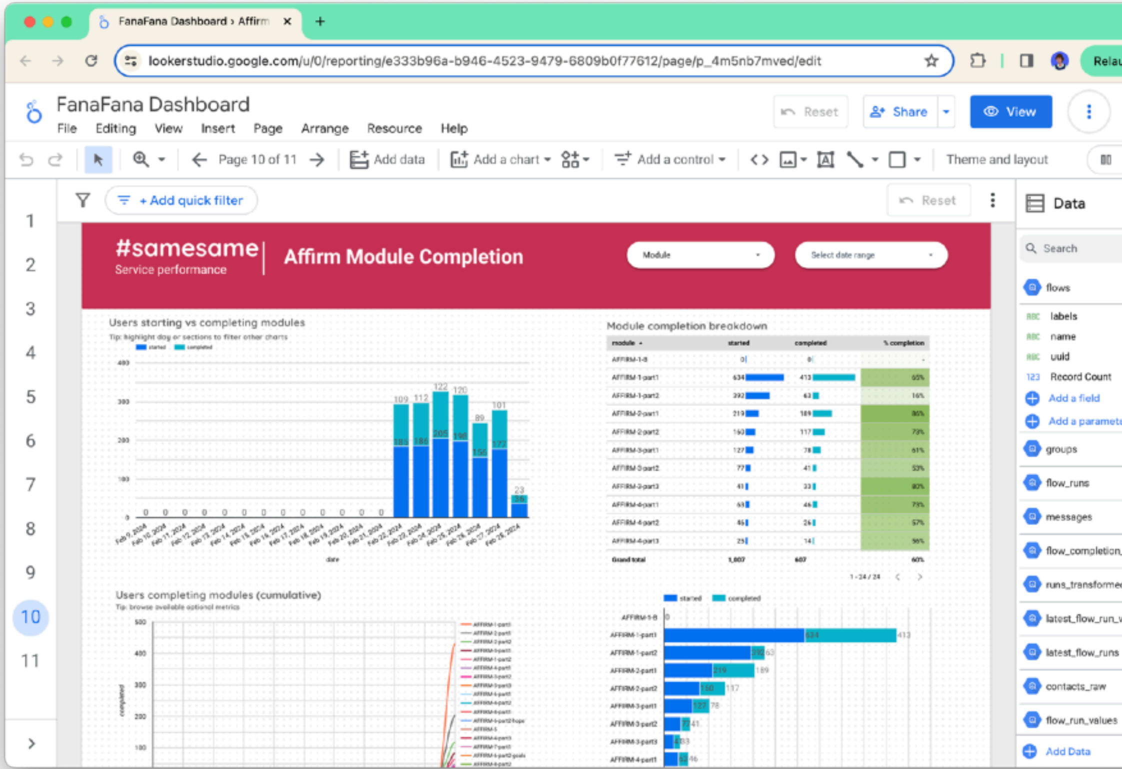
By forcing users to follow a linear path we are seeing better completion rates — more than 2x

Comparing menu options which used buttons or text inputs or menus or a combination

We reduced menu errors (user responses generating a “I don’t understand message”) to zero

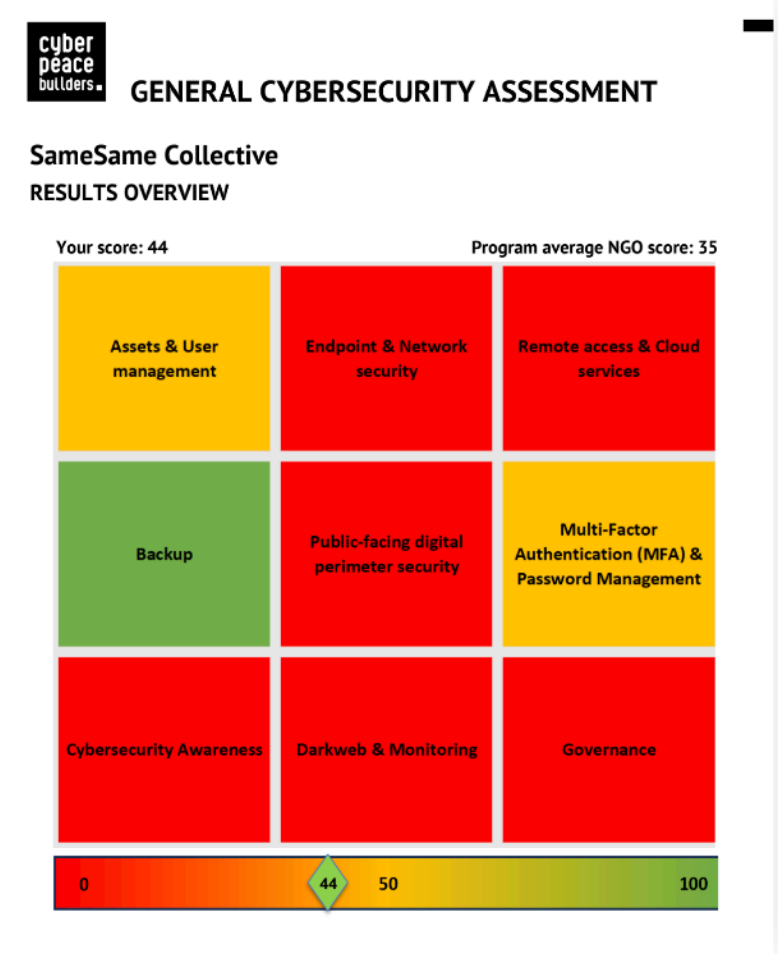
# 02 OUR NEW DASHBOARD

Working with a data scientist, we were able to conduct an audit of our LookerStudio dashboard and, in a series of workshops, to identify and prioritise the metrics we believed were important to track and the ways our conversational flows in TextIt needed to be tagged to ensure we were able to generate these metrics. Our new dashboard has significantly reduced the amount of time that is spent manually analysing data and has increased our confidence in the data we have.



# 03 DATA SECURITY AUDIT

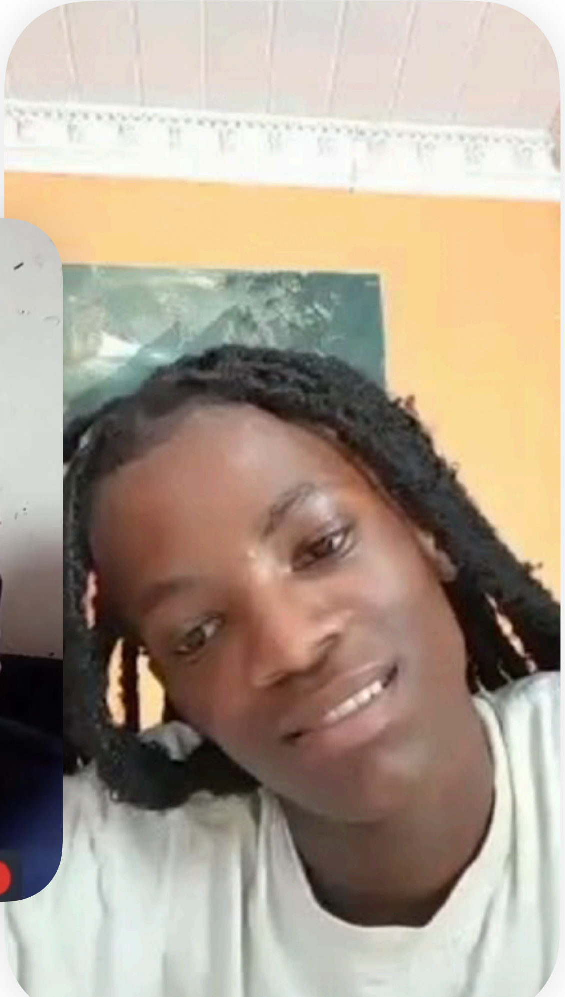
The team at PJMF was able to introduce us to the CyberPeace Builders team who helped us conduct a rapid assessment of our cybersecurity infrastructure. Through this rapid audit we were able to identify several areas that required attention and have either actioned these issues or have plans in place to address them. We were also successful in our application to join the CyberPeace Builder programme and are leveraging their expertise to address these issues. We already have a dark web expert poised to help us, for example.




SAMESAME

# 04 USER RESEARCH

The data we gathered through the experiments we ran through the RL system provided us with valuable insights into what users were doing and how we could improve user engagement and through that, improve their mental wellbeing. In order to understand WHY they were taking specific actions, we conducted user research to try and come to a better understanding of the factors determining users' choices. We conducted a diary study in which we asked 18 queer youth (18-24) to interact with our chatbot.



SAMESAME

# 04 USER RESEARCH FINDINGS

This slide presents a selection of the findings of our user research that are relevant to our RL work:

## 01

### Menu naming

One of the things we found through our experiments is that variations in the number of items in the main menu didn't seem to have any dramatic impact on user engagement. In conversations with users we discovered that what governed their choices with the menu was not the length or complexity of menu items but the extent to which the menu items reflected the CONTENT of the items. We now have a new RL experiment with revised menu options that respond to this finding

## 02

### Length vs content vs speed

Through our RL experiments we explored the extent to which the length and content of messages influenced the likelihood of users continuing to engage with the bot, but one of the findings of our interviews with users was that the SPEED at which multiple messages arrived proved to be an irritation that deterred some (but not all) users to continue. A new RL experiment is being prepared that would test ways to reduce the speed of messages (longer, less frequent messages vs user-triggered) message delivery.

## 03

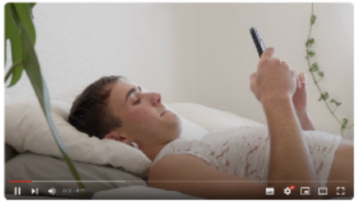
### Prepping users for what to expect

One of the findings of the sessions with users seemed to directly contradict the findings of one of our RL experiments. The leanest onboarding experience consistently delivers better results, but one of the consistent complaints of users we interviewed was that they weren't given ENOUGH INFORMATION about what the bot was about in the onboarding process. This has prompted us to re-examine the goals set for the RL algorithm to take into account completion of multiple 'flows' rather than just the onboarding flow.

# 04 USER RESEARCH

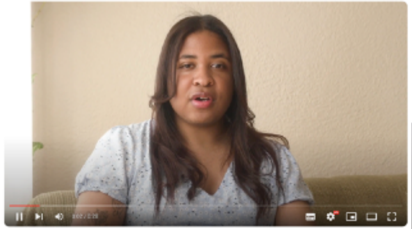
The user research was also used to generate content that supported the creation of new user acquisition campaigns on Facebook and Instagram. We asked our research participants (only once the research had been completed) whether or not they'd be willing to share their experience of using the chatbot with others and asked them to write their own scripts talking about their experiences. Four of the videos produced are available to watch at the links provided.

[Watch John's story]



[Watch Tre's story]

[Watch Orrin's story]




[Watch Zuko's story]

# INSIGHTS

SAMESAME

# INSIGHTS AND LESSONS

Over the course of our participation in the Data & Society Accelerator, we've identified a number of insights related to our efforts to incorporate Reinforcement Learning into our work.

**01**

### Learning questions

Deciding what variations are meaningful to test is as important as building the infrastructure that supports ongoing experimentation. We now have a more intentional approach to learning.

**02**

### Content production

Creating the infrastructure for the experiments wasn't nearly as time-consuming as creating the content needed to make those experiments operational. We now have a much better gauge of the level of effort required to develop content.

**03**

### Mental shifts

Deciding to adopt an approach to learning that prioritises *continuous* learning has required a mental shit on our part — we are never 'done'. Experiments are seldom concluded, even if the data reaches a point of near certainty and while that's required more time managing increasingly complex flows in the back-end, the mental shift has almost been a bigger challenge.

SAMESAME

Patrick J McGovern FOUNDATION

# INSIGHTS AND LESSONS

Over the course of our participation in the Data & Society Accelerator, we've identified a number of insights related to our efforts to incorporate Reinforcement Learning into our work.

## 04

### Premature decisions

We've discovered that a common challenge in the realm of multi-armed bandit experiments, particularly when using algorithms like Thompson Sampling, is that the algorithms can prematurely converge to a particular arm or choice without sufficient sampling. Our solution has been to restart the RL system to validate the findings of initial decisions.

## 05

### A place for humans

Through the discovery that the RL system, without tweaking, will sometimes reach premature decisions, we discovered that human reviews of the data generated by our experiments is still valuable. Similarly, the user interviews we conducted surfaced challenges and opportunities for improvement that were not self-evident through an examination of the data. **Our conclusion is that machine learning systems are not an effective replacement for human decision-making and oversight, but a complement and that the future of SameSame will be determined by our ability to successfully balance both human- and machine-driven decision-making.**

SAMESAME

Patrick J McGovern FOUNDATION

# RISK REVIEW

SAMESAME

# RISK MITIGATION REVIEW

| IMPACT | PROBABILITY | RISK | MITIGATION PLAN | RESULTS |
|---|---|---|---|---|
| HIGH | LOW | Meta's platforms are/ don't prove to be an efficient means of recruiting users | We will partner with on-the-ground organisations to leverage their peer mobiliser networks to promote the bot | Meta's platforms have proven to be an efficient means of recruiting users, but we've supplemented these recruitment efforts with partnerships with zero-rated platforms (like the Moya platform in South Africa), that have proven to be even more effective at user acquisition, |
| MED | MED | It takes too long or is too complicated to produce the variations in content required to test the system | We will first test the system with the simplest content sets (long vs short messages) to make sure the system works at all before adding in more variations. | We correctly identified this as a risk in the early stages of the project as the delays in content production contributed to delays in the close of the project. That said, it was deciding WHAT to test and then developing the content plans |
| HIGH | LOW | The number of users is too small for the system to learn anything conclusively or the difference in performance between the different versions is too small for the RL system to make effective decisions. | If user acquisition is a problem, we will divert funds away from content creation and test fewer variations to move more budget into user acquisition. | Ultimately, we've been able to recruit large numbers of users — sufficient numbers to achieve statistical significance. What we did not appreciate is that even with large numbers of users that the RL system still might need some human oversight to make good decisions and avoid premature ones |
| HIGH | MED | The platforms/ tech stack we have chosen to host early versions of the bot are outgrown quickly or are not sufficient to satisfy org. needs | We are actively partnering with tech experts to help us assess our current tools to assess for short/ med./ and long-term usability. We are developing budgets to ensure tech stack upgrades or changes are possible when needed. | We continue to assess the options available to use as new platforms and features are released, but nothing has emerged to date that has convinced us of any urgent need to change up our tech stack, |

# THANKS
# NDATENDA
# NGIYABONGA

SAMESAME