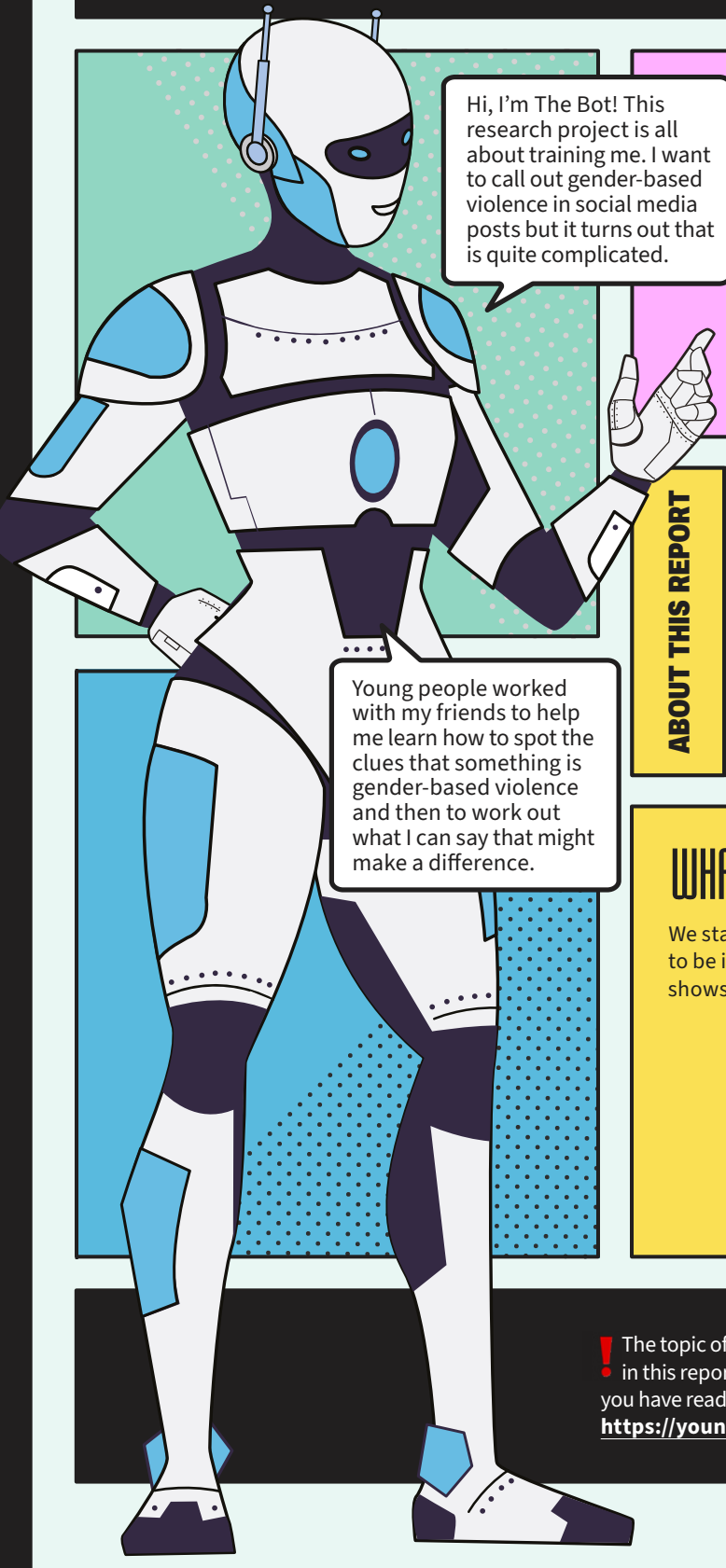
EQUALLY SAFE ONLINE

YOUNG PEOPLE VS. GENDER-BASED VIOLENCE



ABOUT THE RESEARCH PROJECT

Equally Safe Online (ESO) is a research project exploring how artificial intelligence (AI) can help people who are experiencing online **Gender-Based Violence (GBV),** mainly looking at social media platforms. The project was funded by the **Engineering and Physical Science Research Council (EPSRC)** and the researchers are from **Heriot-Watt University** and **The University of Edinburgh.** We also worked closely with organisations helping people subject to GBV.

ABOUT THIS REPORT

Researchers, young people and an **artist** worked together to create this graphic report to share important findings from our workshops with **24 young people aged 13-20.** The report explores young people's ideas about how to train AI to detect online GBV and to identify what kinds of responses the AI could make that would be most effective. It also shares how these ideas have been used by Computer Scientists to help build The Bot.

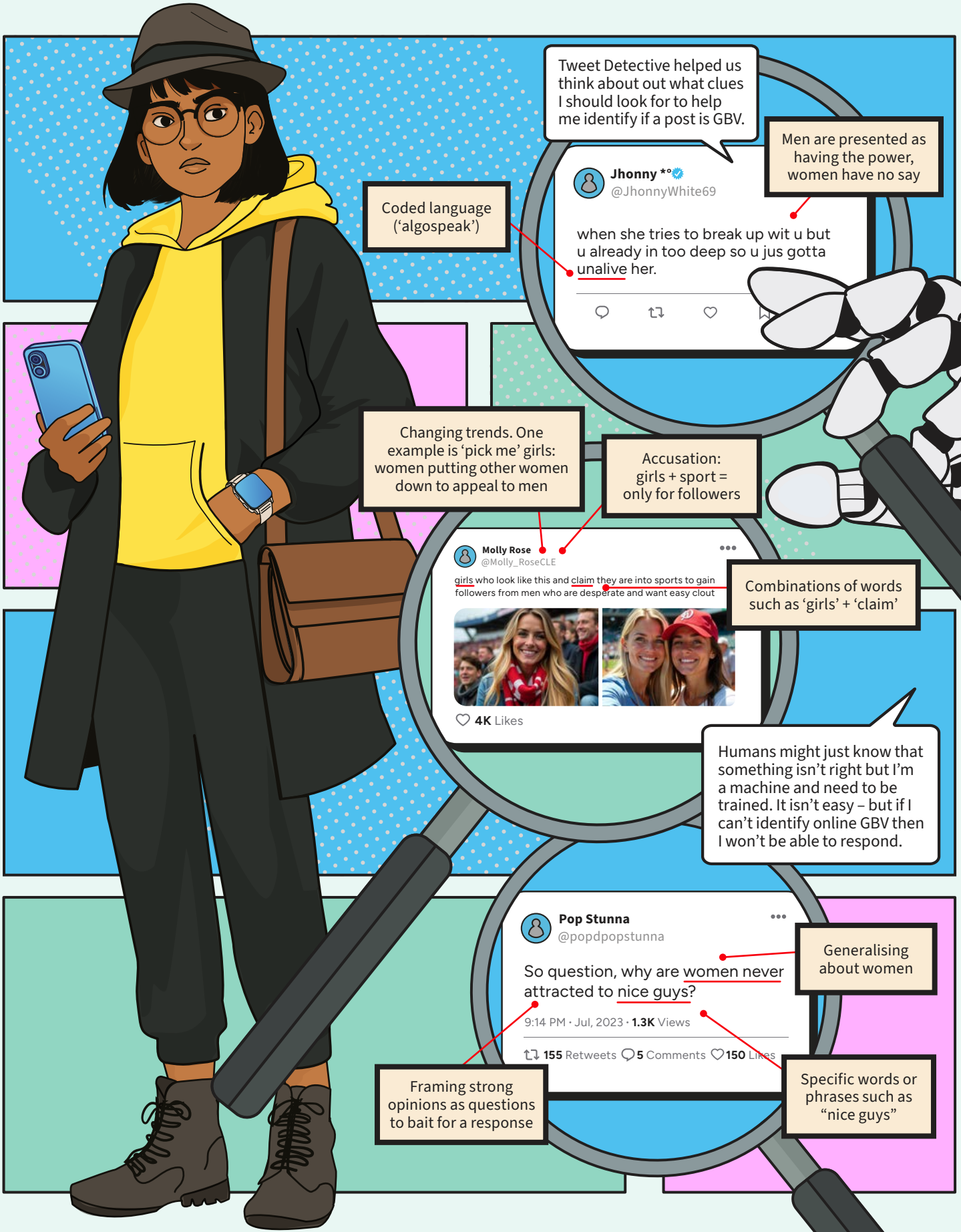
WHAT IS GENDER-BASED VIOLENCE?

We started each workshop by asking participants what they understood to be included in the phrase **'Gender-Based Violence'**. This word-cloud shows all the responses. The bigger the word, the more people said it:

Sexual violence
Rape
violence_bc_of_gender
Threat
minorities
abuse
inequality
right_wing_views
exploitation
Violation_of_Human_Rights
Domestic violence behaviour
discrimination
Coercive_control
gendered_violence
harrassment
Controlling behaviour

! The topic of this research is difficult and we understand that there may be things in this report that readers might find upsetting. If you have been upset by anything you have read please talk to a trusted adult or find information/get support here: **https://young.scot/get-informed/thatsnotok-support-information**

TWEET DETECTIVE CLUES



TWEET DETECTIVE CONTEXT

Avatar details

Username

Bluetick status: trust more? Or paid for?

Fake user 69 AB @fake_user69ab

Taylor Swift is a lunatic.
She should just stay at home.
Her music sucks.

500k Retweets **67k** Comments **2M** Likes

A.I. generated imagery

Tom Husting @tomhusting

Follow

Joined December 2019

2.5K Following **482.3K** Followers

Pinned Tweet

Tom Husting @tomhusting

Real women should be submissive and not argue. Do what you're told. Make yourself pretty.

102 Retweets **20** Comments **421** Likes

Lots of different reactions – controversial?

This one's gone **viral**

Telling women what they should do

Happy Tuesday!

Fake user 69 AB @fake_user69ab · 1hr

Check her out 😈

John Smith @tatefan_1978 · 1hr

I'd still 🍇 her

GoodBadUgly @xulu175839BV · 1hr

🤙

Raina Watkins @rainshine_15 · 1hr

That isn't something you should joke about

Turler Dardin @fitclarb_x_28

The feminazis are out in force in these comments

Gamers Delight @robloxforever2011

These nasty feminist bitches sho… can't take a joke

Sometimes the clues aren't just about the words in the post. It makes a difference who posts it and who reads it and what else they say.

TWEET DETECTIVE WHAT WE DID

SENSE-CHECKING OUR FLAGGING SYSTEM

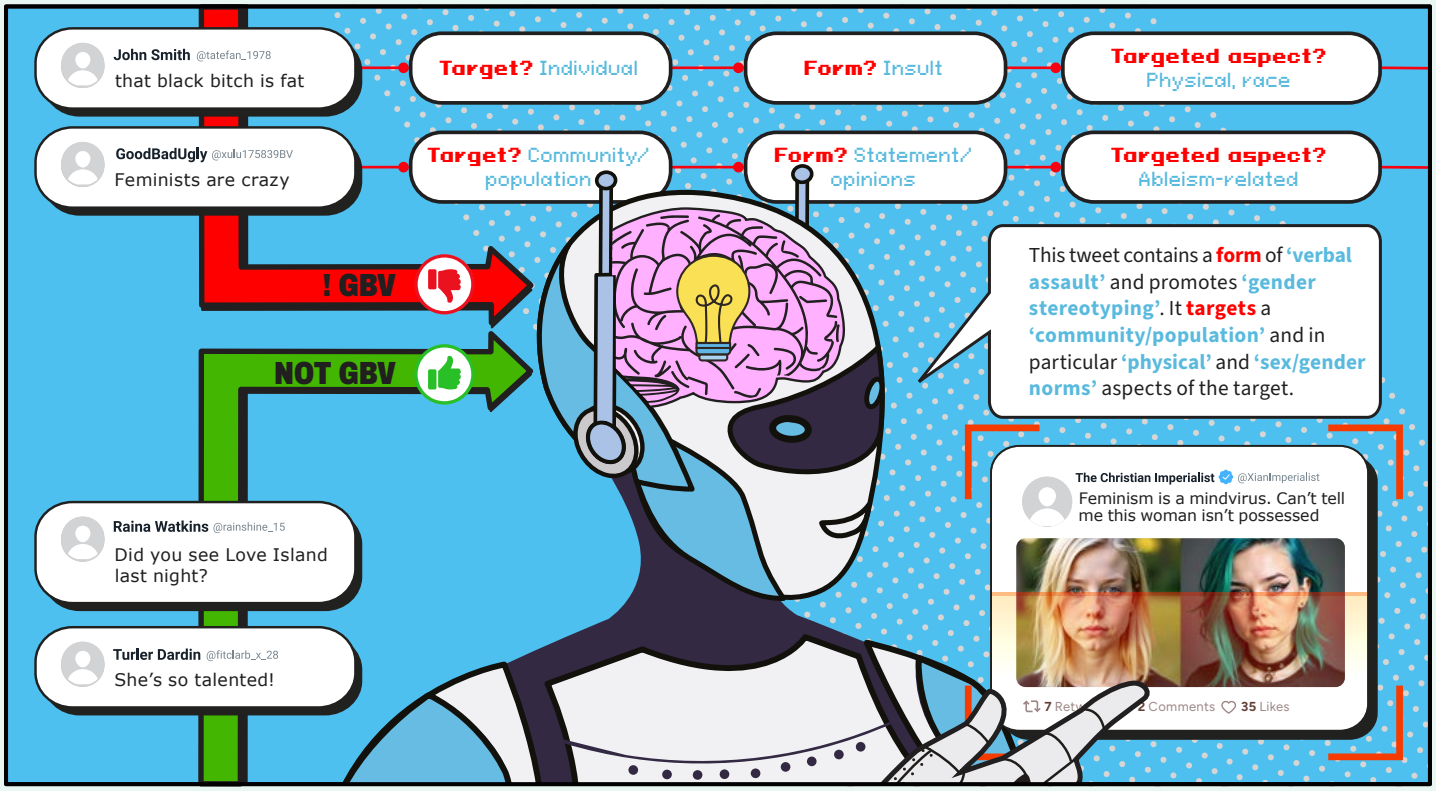
Young people helped the researchers understand what online gender-based violence (online GBV) looks like for them. This helped us "sense check" what adult experts had told us earlier in the project. The adults created a flagging system for Online GBV, which we use to train The Bot.

Before, we only used one flag, e.g. "is it toxic language, or not?". But after speaking to the adults, we now have a whole set that describe the online GBV in detail. We can use the extra information to create better responses.

It was important to make sure the flagging system reflects young people's experience online. And the good news is that it does! Young people used similar language and described similar things happening online to the adults.

From what they said, it was clear young people understood complex-sounding concepts like violation of gender norms, intersectionality, and internalised misogyny, all of which are captured in what we call the "taxonomy."

The visual below shows how The Bot learns.



IDENTIFYING CLUES

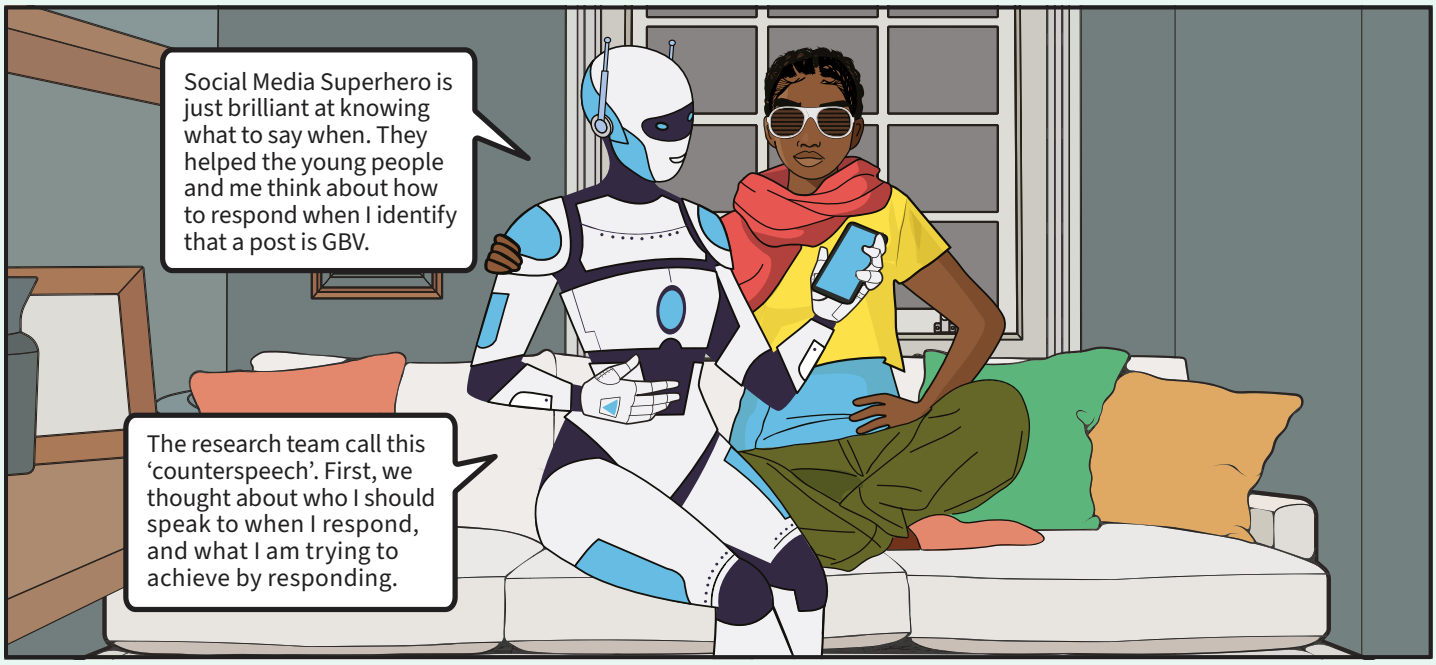
Young people helped researchers identify the clues to look for to judge if a post is online GBV. This helped us choose which social media posts to collect to train the first version of 'The Bot'.

- For example, we used key "alarm bell" phrases and trends such as "Show me a female" and #YourBodyMyChoice as search terms when collecting the data.
- Based on our discussions with young people we also included images, memes and emojis in the data collection. Often, AI research on online GBV just focuses on text (words).
- We also collected 'meta data' for each post; details of who is posting, the number of likes, re-posts and replies. This is also a new approach in AI research on online GBV. We have collected **> 7,000** social media posts (in their conversational context) for training The Bot.

Young people helped the researchers understand what online gender-based violence (online GBV) looks like for them. They also helped researchers identify the clues to look for to judge if a post is online GBV.

SOCIAL MEDIA SUPERHERO

COUNTERSPEECH AUDIENCE

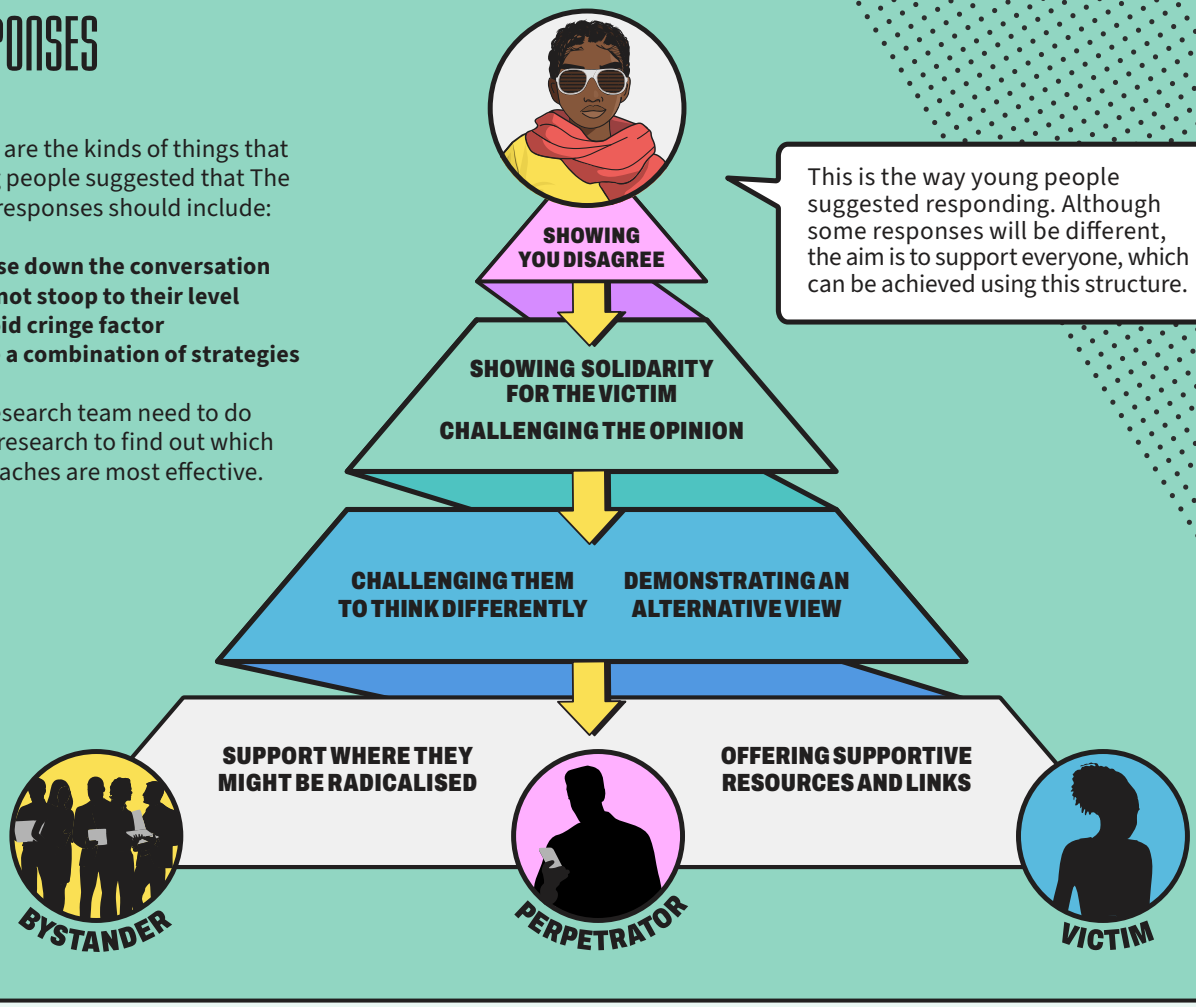


RESPONSES

These are the kinds of things that young people suggested that The Bot's responses should include:

- **Close down the conversation**
- **Do not stoop to their level**
- **Avoid cringe factor**
- **Use a combination of strategies**

The research team need to do more research to find out which approaches are most effective.



This is the way young people suggested responding. Although some responses will be different, the aim is to support everyone, which can be achieved using this structure.

SOCIAL MEDIA SUPERHERO COUNTERSPEECH STRATEGIES

EXAMPLES OF POSTERS THE BOT MIGHT NEED TO RESPOND TO:

HIGH PROFILE MIDDLE-AGED MALE INFLUENCER

He makes money from feeding off young men's insecurities so The Bot isn't going to change his mind. Using humour or being sassy might be good here and also telling his supporters the facts.

EXAMPLE RESPONSE:

"This guy's wanted for trafficking, maybe don't take his advice."

TEENAGE BOY

He might not know much about what he is posting or why it isn't okay. The Bot needs to educate him without pushing him further down a radicalised rabbit-hole. It might be useful to point out the possible consequences of his post.

EXAMPLE RESPONSE:

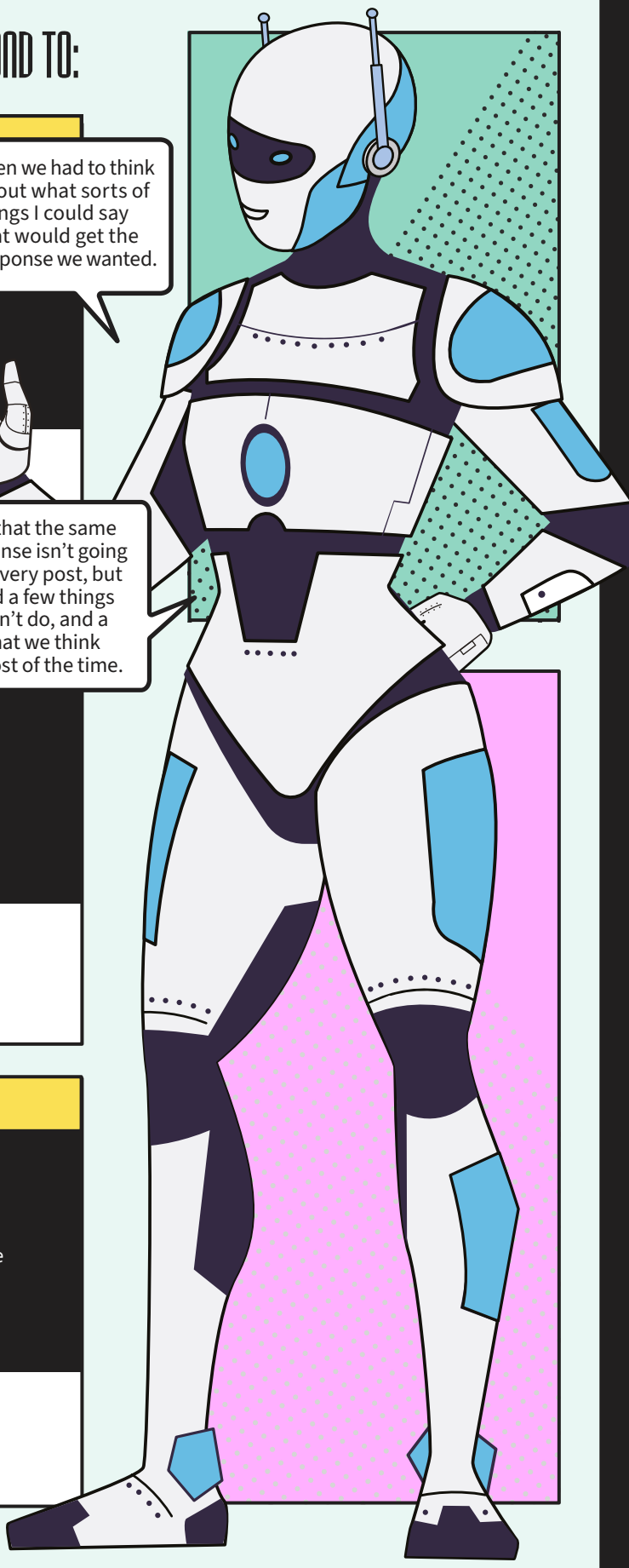
"Your comments contribute to hurting other people. For more info, go to [link]."

YOUNG FEMALE SOCIAL MEDIA INFLUENCER

She colludes with GBV by telling women how they should live their lives. The Bot should point out the hypocrisy and affirm that different women should be able to choose how they want to live their lives.

EXAMPLE RESPONSE:

"It's okay for everyone to not want the same thing."



SOCIAL MEDIA SUPERHERO WHAT WE DID

WHAT WE DID: COUNTERSPEECH

Young people's input is helping researchers understand how to respond to posts that have been flagged as gender-based violence.

1. We used what young people told us to identify the gaps in existing AI research on this topic, so we can highlight them to the community and provide good-practice guidelines.

2. We've also experimented with training AI models using the online GBV's features, e.g. who it's aimed at, whether it takes the form of a joke or a threat etc., to generate counterspeech.

   This includes an experiment where we instructed the model to use different combinations of strategies (including facts + humour!). Other research has focused on using one strategy at a time.

THE RESULTS

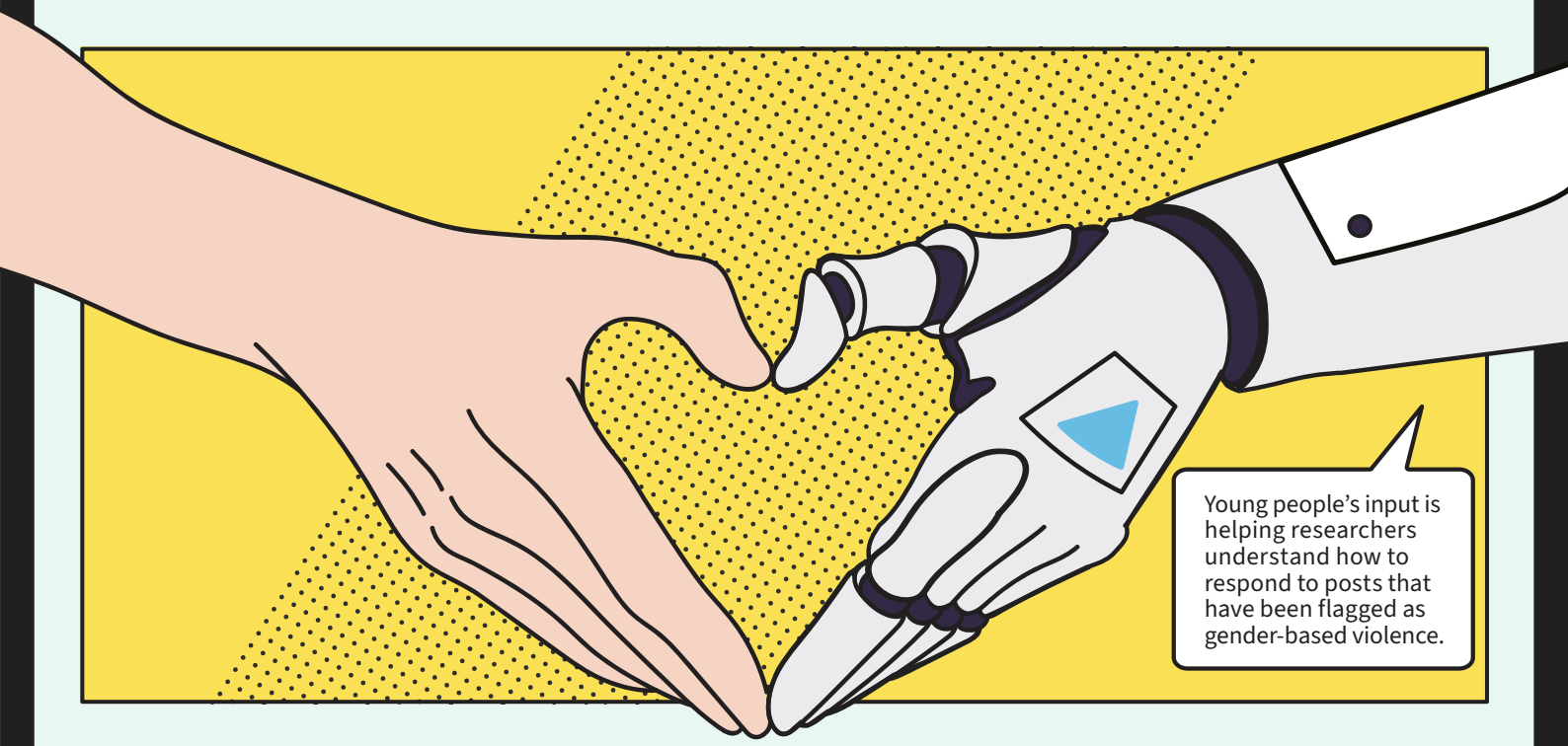
Results are promising! They show AI-generated counterspeech that combines multiple strategies or is aware of the online GBV's features - or both - is more persuasive and more educational than counterspeech written by humans!

This is exciting because generating counterspeech is even more difficult than detecting online GBV. There is no simple solution but the young people in this project have already made a significant contribution to AI research, including in setting the direction for future work.

WHAT WE'RE DOING NEXT

The hard work everyone has done is being put to practical use in a new project at Heriot-Watt, which will build a web-based 'Support Buddy' to detect online GBV & take action against it. The plan is to make this available for anyone to download. Keep an eye on our website for updates!

Young people recommend co-creating educational materials for schools based on the workshops and what we've learned.



YOUNG PEOPLE KEY MESSAGES

"I want people to be aware of how far people go online and how **extremist** their points of view are."

—Sana, 17

"I want the reader to know that this happens to **everyday people** all the time and it has real impact on their lives. This research creates awareness of the effects of GBV."

—Angel, 16

"I hope The Bot can do the work of **protecting people online.** This saves others from becoming involved in comments around GBV."

—Bea, 19

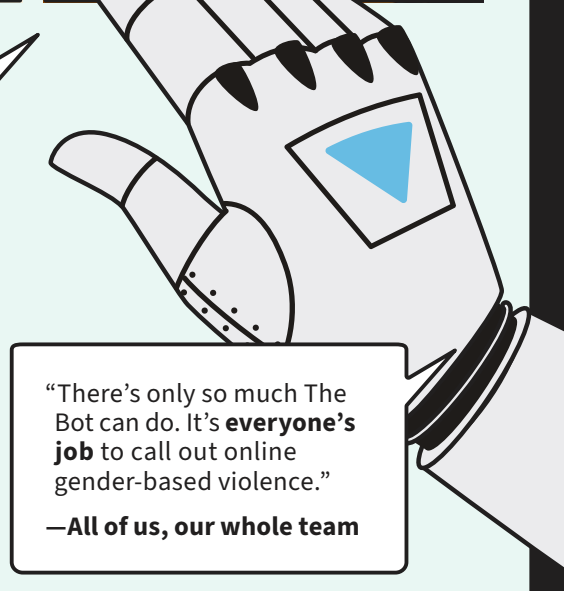
CREATING THIS REPORT

"I want The Bot to create awareness of online GBV and call for action to **help and support** those affected whilst creating safe spaces for online collaboration."

—Laila, 16

"We like that the process was educational and we hope that **young people can continue to help** computer scientists and researchers to train The Bot."

—Honey, 17



"There's only so much The Bot can do. It's **everyone's job** to call out online gender-based violence."

—All of us, our whole team

COLLABORATION

This visual report was developed in a day-long workshop with **5 young people** who all took part in focus groups, **two researchers, two support workers** and an **artist.** We worked together to think about how best to share our research findings in a way that makes them interesting and accessible for as many people as possible. For more information about **Equally Safe Online** please go to: **https://sites.google.com/view/equallysafeonline**

! The topic of this research is difficult and we understand that there may be things in this report that readers might find upsetting. If you have been upset by anything you have read please talk to a trusted adult or find information/ get support here: **https://young.scot/get-informed/thatsnotok-support-information**

HERIOT WATT UNIVERSITY

THE UNIVERSITY of EDINBURGH