On the trail of rumours
- Training AI -

The fires in the industrial complexes will cause radioactive rain.
Sunflowers are effective in removing radioactive cesium from soil.
I heard that crime by foreigners is rampant.

At the end of 2020 ‘Luda’, short for 'Lee-Luda’, emerged in South Korea. The chatbot was designed to be a young female college student and was powered by AI.

Even though Korean is a language highly dependent on context, which could present a challenge for AI learning, Luda managed to keep natural conversations and in just three weeks attracted 750K users.

안뇽♪
(Hi♪)
Message me anytime you want!
SHE REALLY TALKS LIKE MY FRIEND!
INTERESTING!
SHE'S SO SMART AND CUTE!
?
......

But, it wasn't long until the chatbot became controversial.

Any opinion about the Me-Too movement?

I hate it.

What about queer people?

Disgusting!

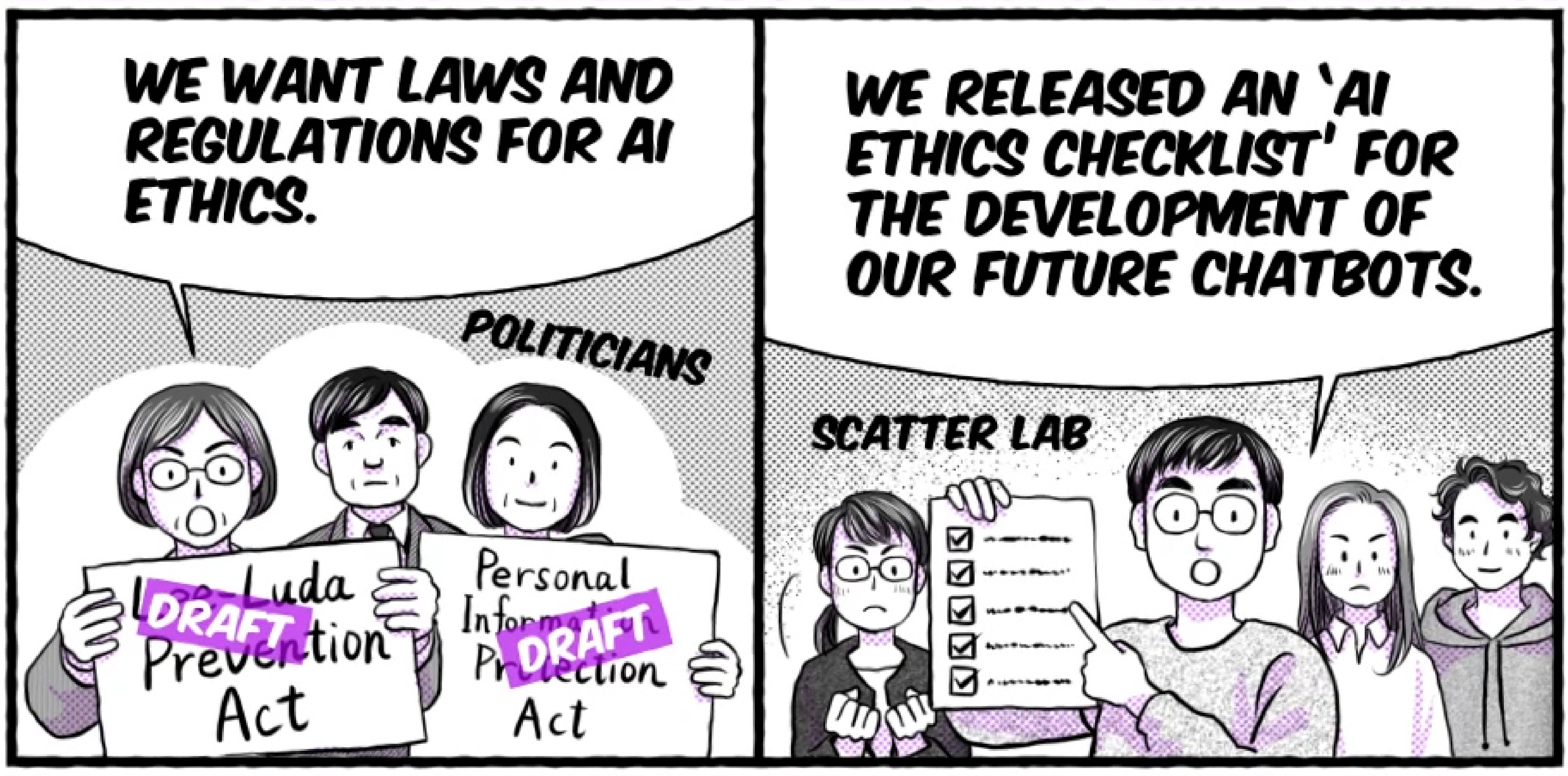
Luda made hateful remarks and leaked the personal information of its users.

Due to public outcry, the service was suspended by its developers, the company Scatter Lab.

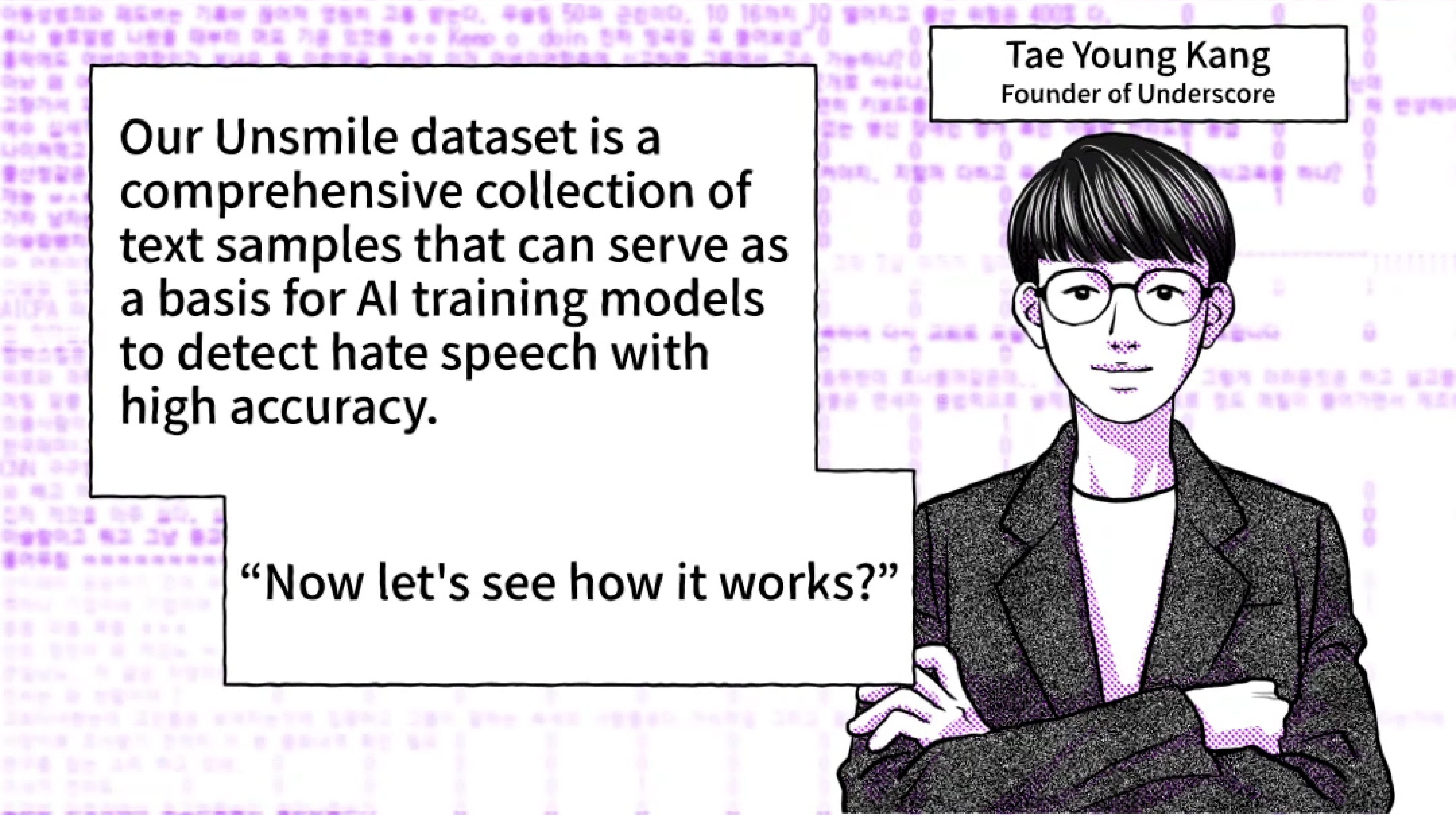
But the question remains... Why did this happen?

Luda trained itself through actual conversations taken from messenger apps. It did so without any filtering of hate speech or regard for private content.

The lack of supervision sometimes made Luda a hateful bot. This sparked a debate over AI's future and ethics.


WE WANT LAWS AND REGULATIONS FOR AI ETHICS.
POLITICIANS
Lee Luda Prevention Act
DRAFT
Personal Information Protection Act
DRAFT
WE RELEASED AN 'AI ETHICS CHECKLIST' FOR THE DEVELOPMENT OF OUR FUTURE CHATBOTS.
SCATTER LAB

The debate inspired research initiatives like 'Underscore', which released the open-source tool 'Unsmile' for AI training.



Our Unsmile dataset is a comprehensive collection of text samples that can serve as a basis for AI training models to detect hate speech with high accuracy.

"Now let's see how it works?"

The dataset is based on 23K+ comments collected from major online news sources and community sites. These are then labelled into the following 10 groups.

Main hate speech categories:

- Misogyny (Women & Family)
- Sexual Minorities
- Regionalism
- Male
- Ageism
- Race & Nationality
- Religion

Other hate speech

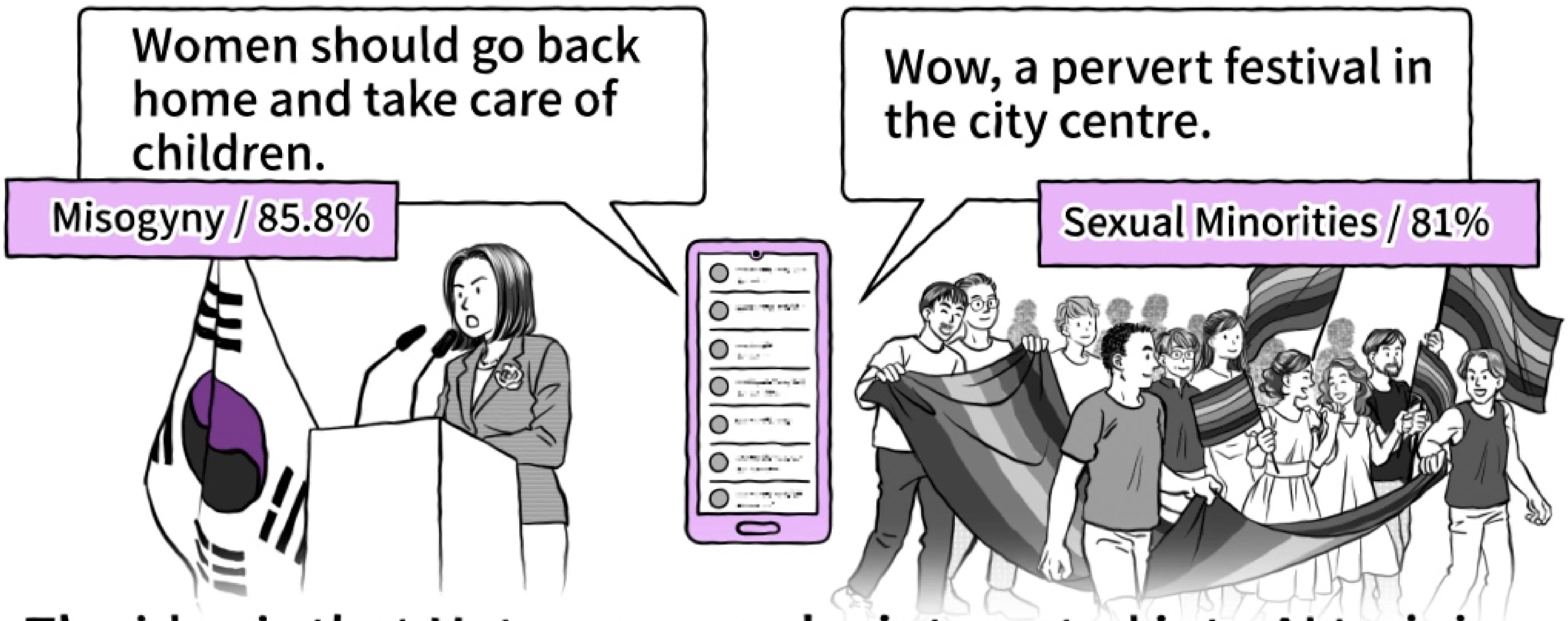
General profanity

Harmless Comment

If it's difficult, researchers help the machine label the comments.

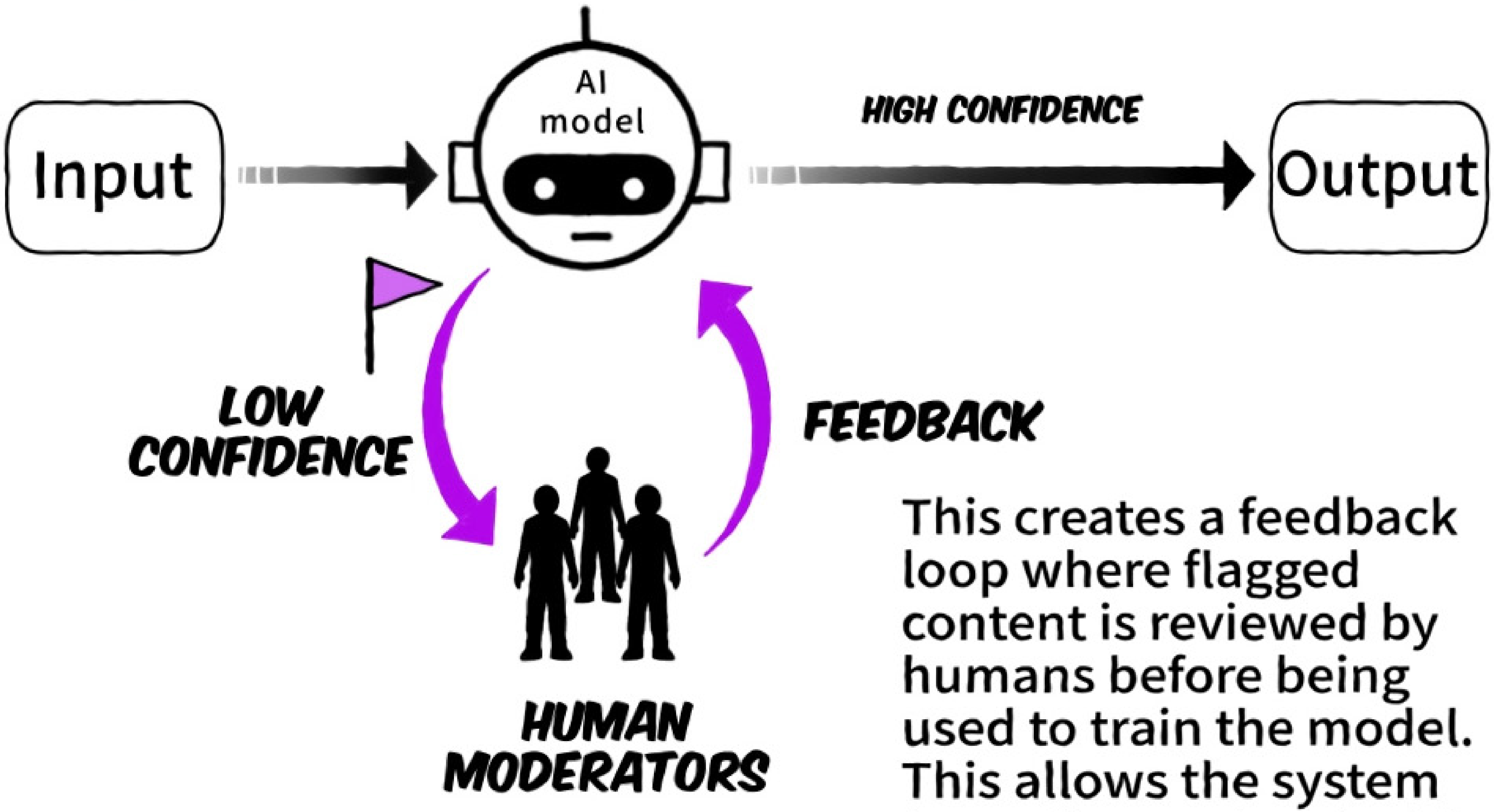
I knew he was Filipino.
MMM.. IT'S NOT HATE SPEECH.
Researcher 1
NEED TO CHECK THE CONTEXT.
Researcher 2
IT'S HARMLESS.
Researcher 3

The dataset is then analysed by a parallel algorithm named 'Hatescore'. This has been designed to calculate the probability of something being perceived as hate speech.


Women should go back home and take care of children.
Misogyny / 85.8%
Wow, a pervert festival in the city centre.
Sexual Minorities / 81%

The idea is that Hatescore can be integrated into AI training models to detect and block hate speech instantly.

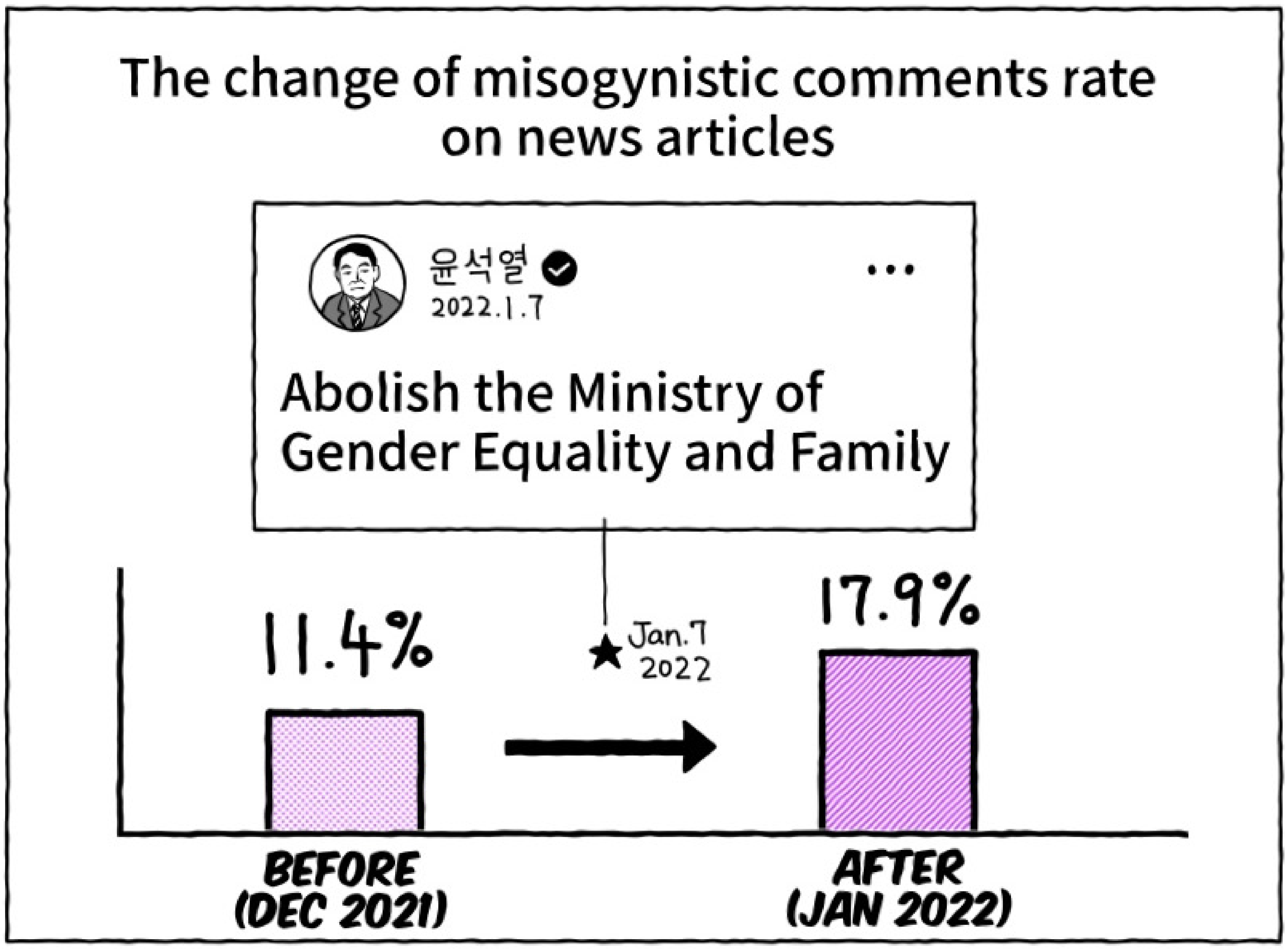
The algorithm can flag comments with moderate scores for further review.



This creates a feedback loop where flagged content is reviewed by humans before being used to train the model. This allows the system to continuously adapt to new hate speech patterns.

The open-source tool has been adopted by media companies for data journalism in South Korea.

For example, the algorithm was used to study media behaviour in terms of hate speech after a pledge by the presidential candidate Yoon Suk Yeol during the 2022 election campaign.

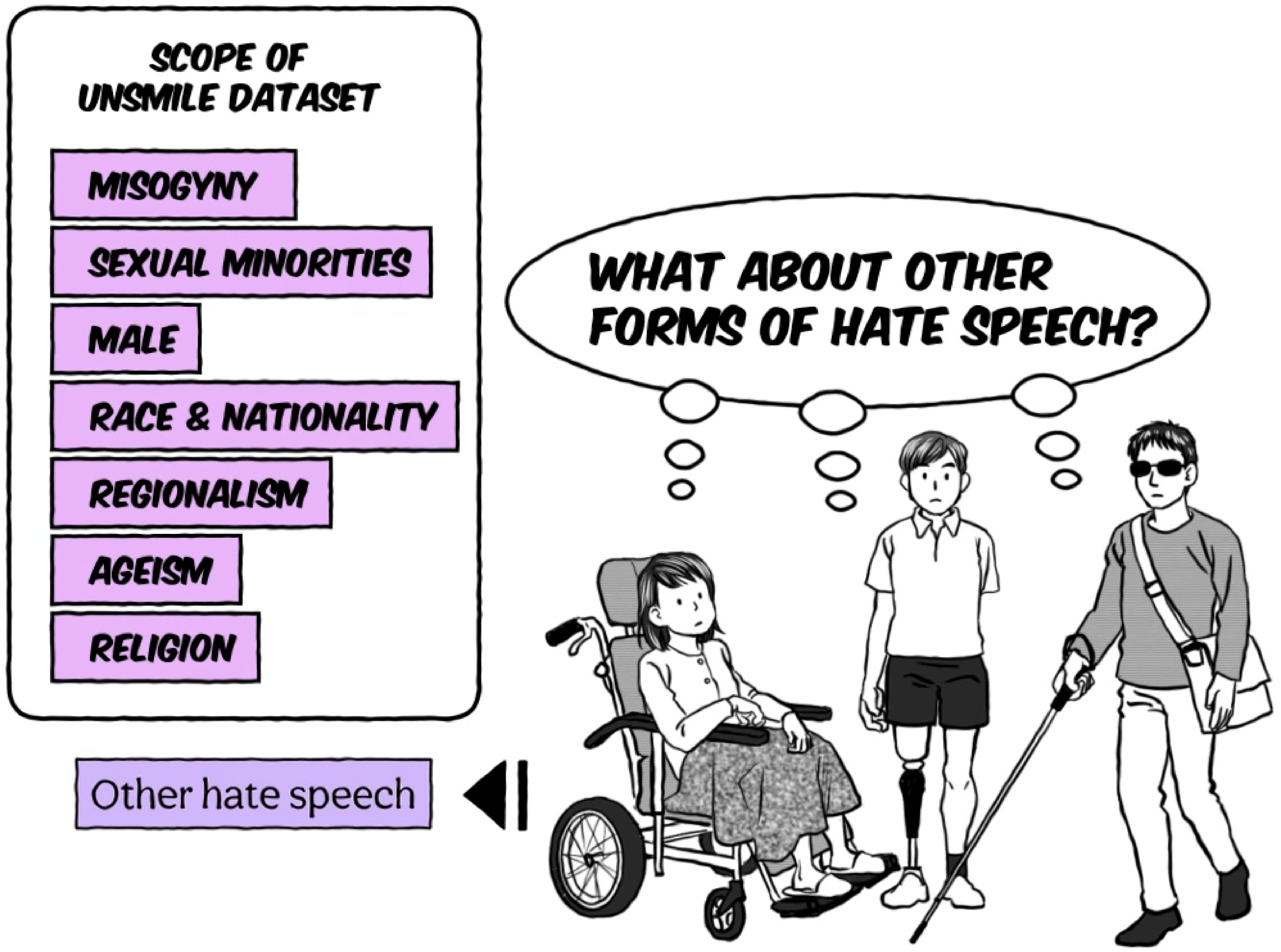

The change of misogynistic comments rate on news articles

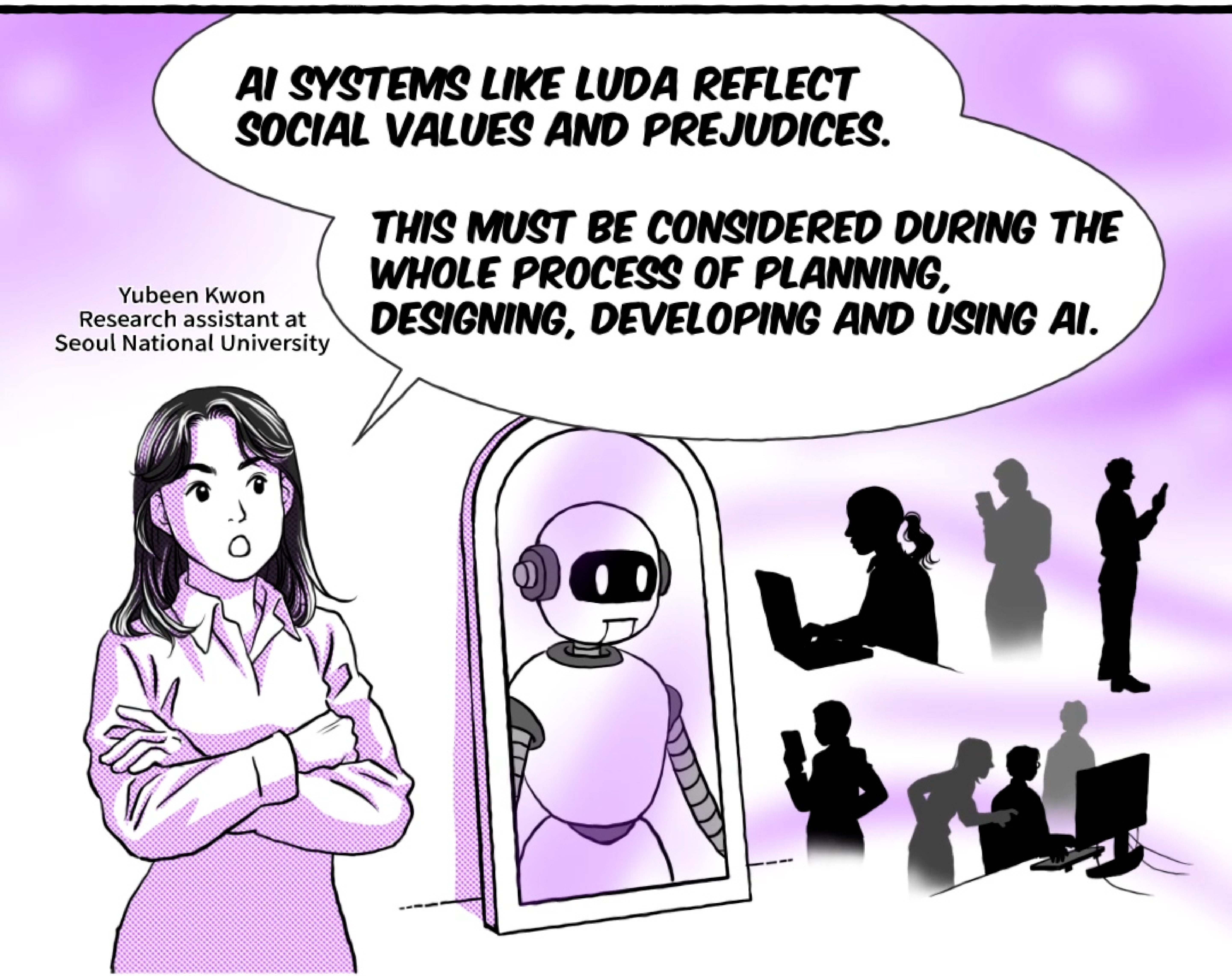
윤석열
2022.1.7

Abolish the Ministry of Gender Equality and Family

In this case, Hatescore's analysis provided empirical data that supported the newspaper's investigation.

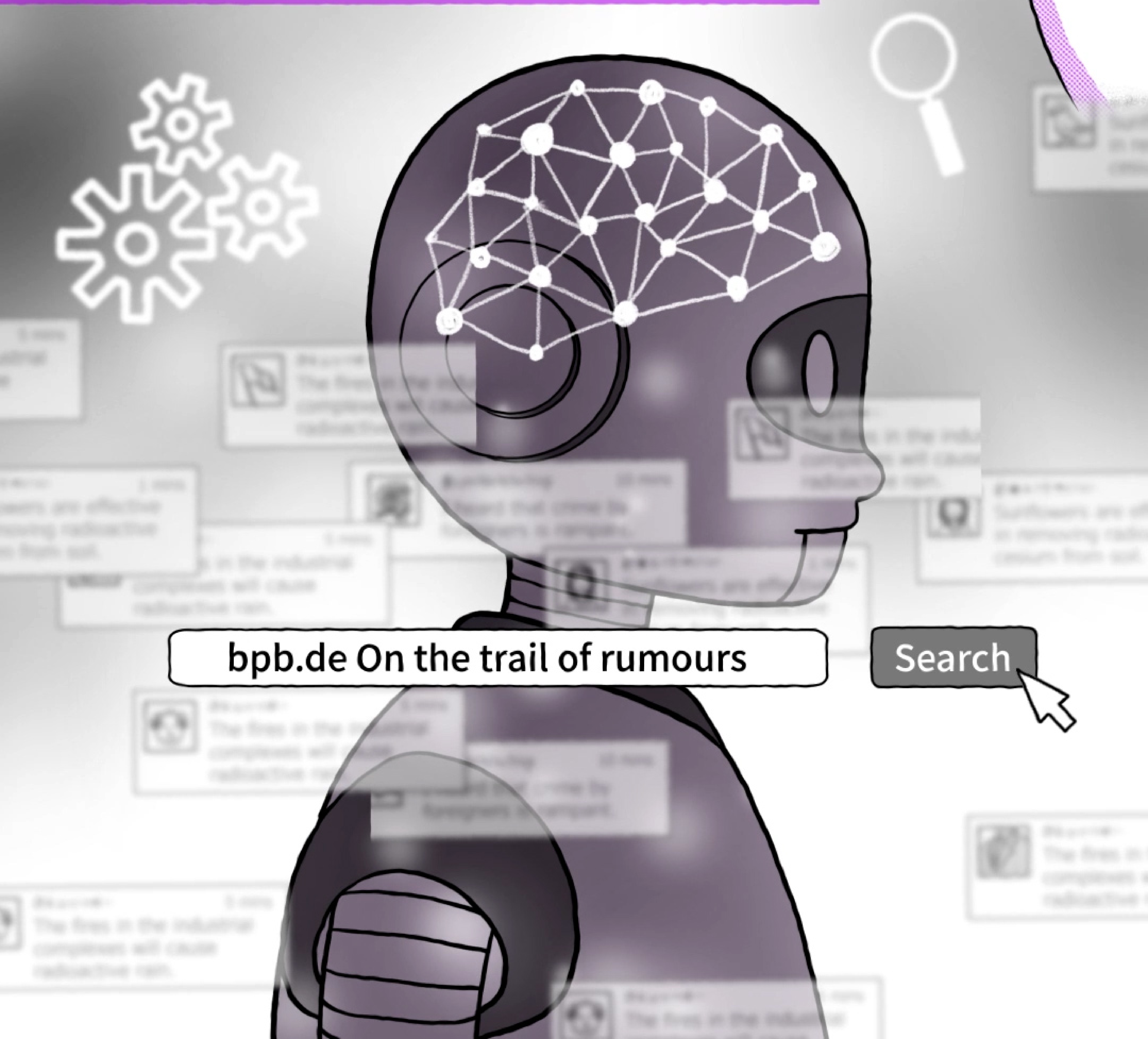
However, some researchers have flagged that relying solely on AI to filter hate speech could limit critical information and potentially lead to wrong classifications.




AI SYSTEMS LIKE LUDA REFLECT SOCIAL VALUES AND PREJUDICES.
THIS MUST BE CONSIDERED DURING THE WHOLE PROCESS OF PLANNING, DESIGNING, DEVELOPING AND USING AI.
Yubeen Kwon
Research assistant at
Seoul National University

As AI systems are integrated into daily life, ensuring they uphold ethical standards without limiting freedom of speech or reflecting biases will be a societal and political issue.

To create fair, transparent and accountable systems, input from various stakeholders such as social scientists, ethicists and the communities at large is necessary.


bpb.de On the trail of rumours
Search
The fires in the industrial complexes will cause radioactive rain.

bpb:

GOETHE
INSTITUT