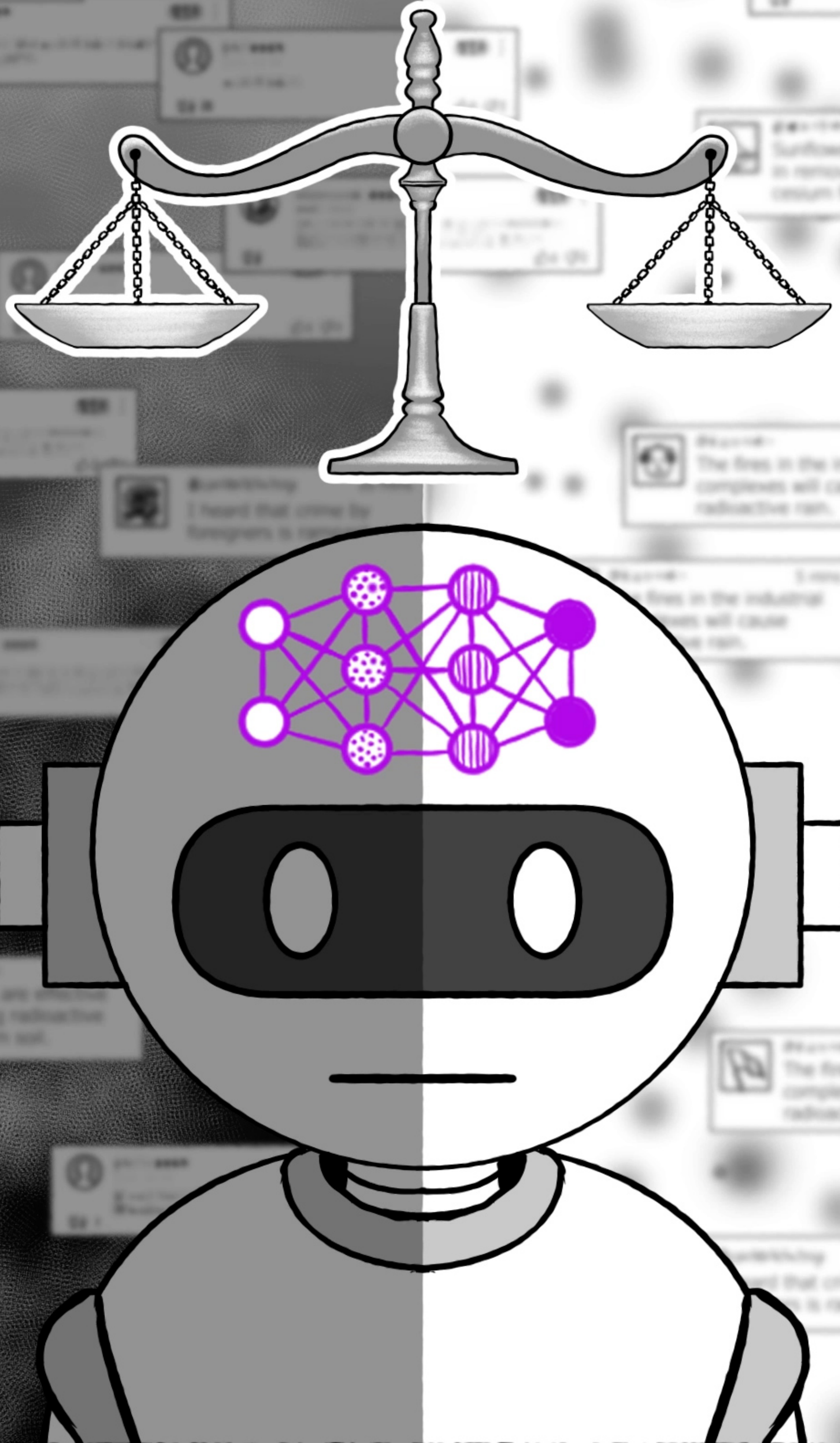
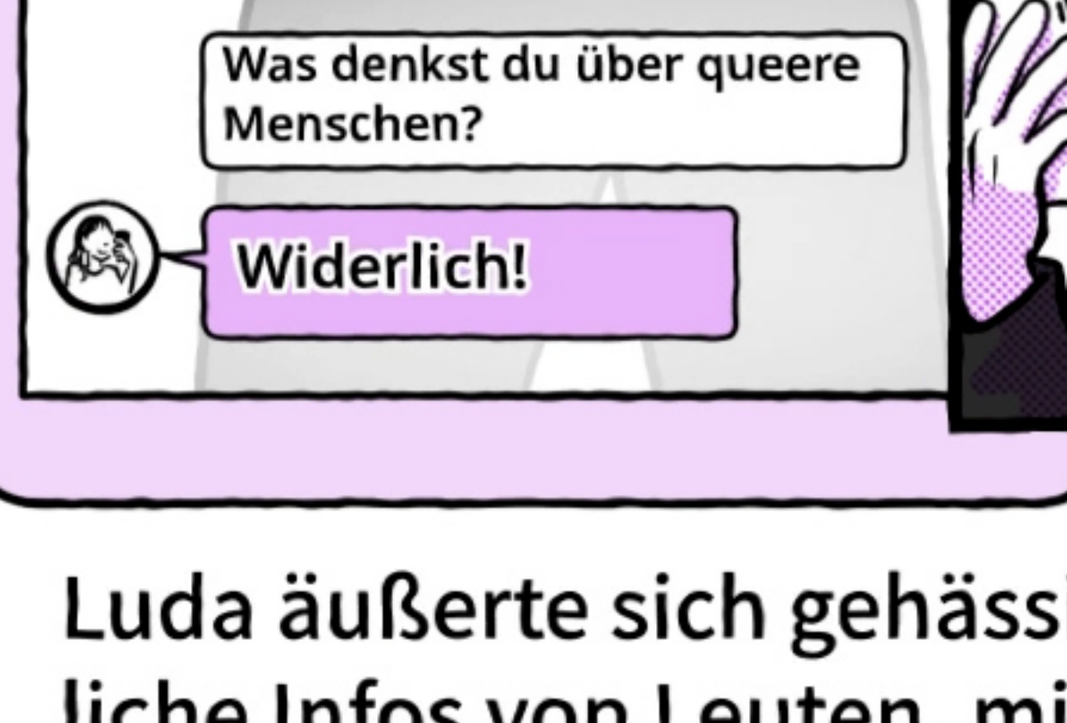
Den Gerüchten auf der Spur
Training für die KI



Ende 2020 tauchte „Luda“, kurz für „Lee-Luda“,
in Südkorea auf – ein KI-betriebener Chatbot,
der eine College-Studentin imitiert.

Koreanisch ist sehr kontextabhängig, was herausfordernd für maschinelles Lernen sein kann. Trotzdem war Luda in der Lage, natürlich zu kommunizieren und hatte nach drei Wochen schon 750.000 Nutzende.

Luda äußerte sich gehässig und verbreitete persönliche Infos von Leuten, mit denen sie sprach.

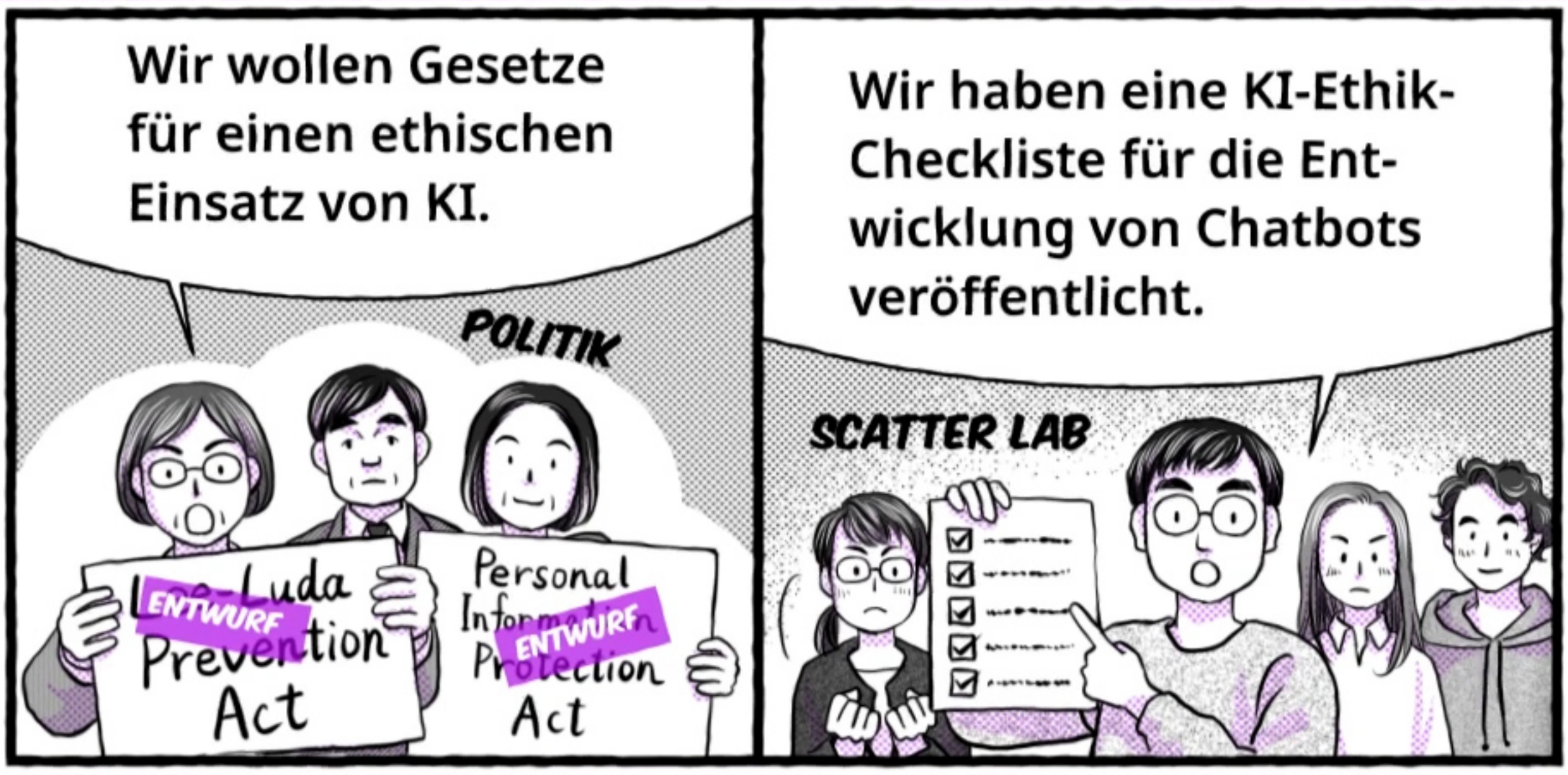


Nach einem öffentlichen Aufschrei nahm das Entwicklungsbüro Scatter Lab den Dienst aus dem Netz.

Doch die Frage bleibt:
Wie konnte das passieren?

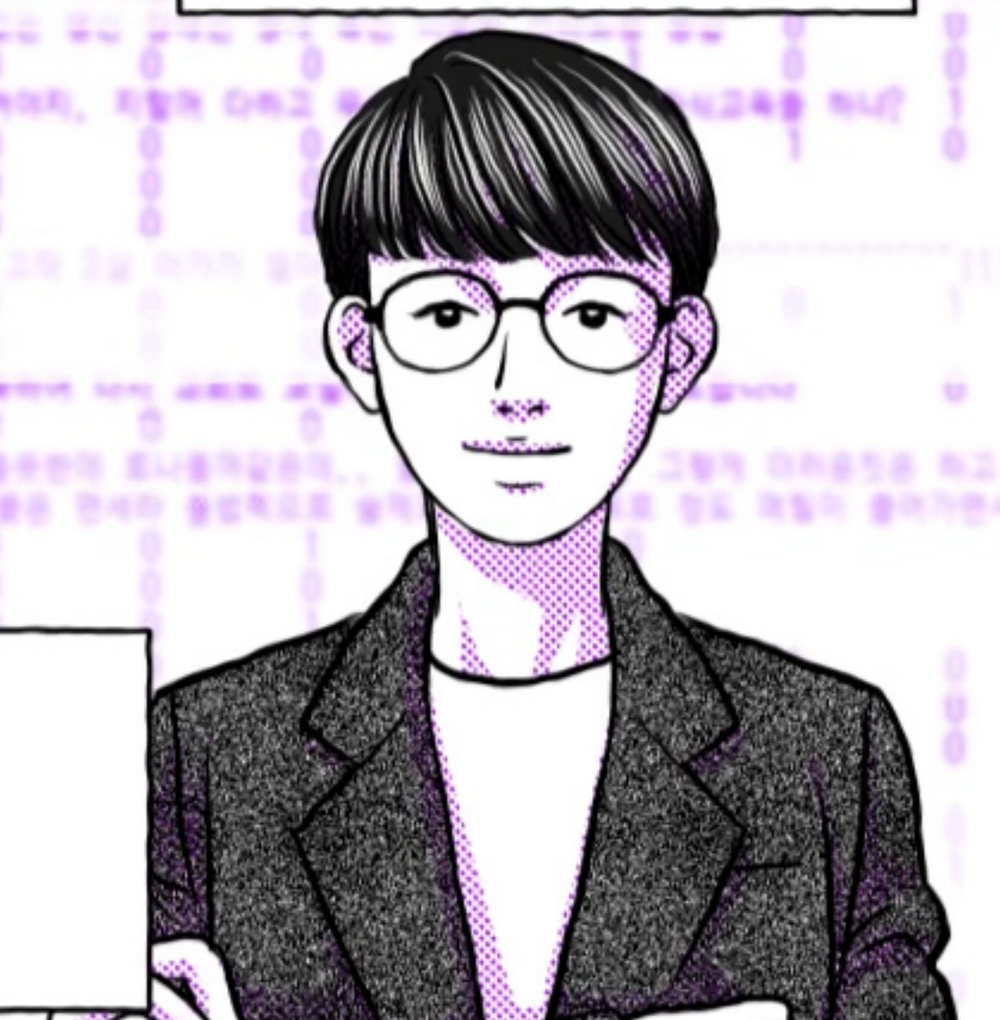
Luda lernte aus echten Unterhaltungen in Messenger-Apps. Dabei wurden weder Hasskommentare herausgefiltert noch auf den Schutz privater Inhalte geachtet.

Durch die fehlende Aufsicht entwickelte sich Luda manchmal zu einem Hass-Bot. Dies löste eine Debatte über Zukunft und ethische Grundsätze von KI aus.



Die Debatte inspirierte Forschungsinitiativen wie „Underscore“, die das Open-Source-Tool „Unsmile“ für KI-Trainings entwickelte.

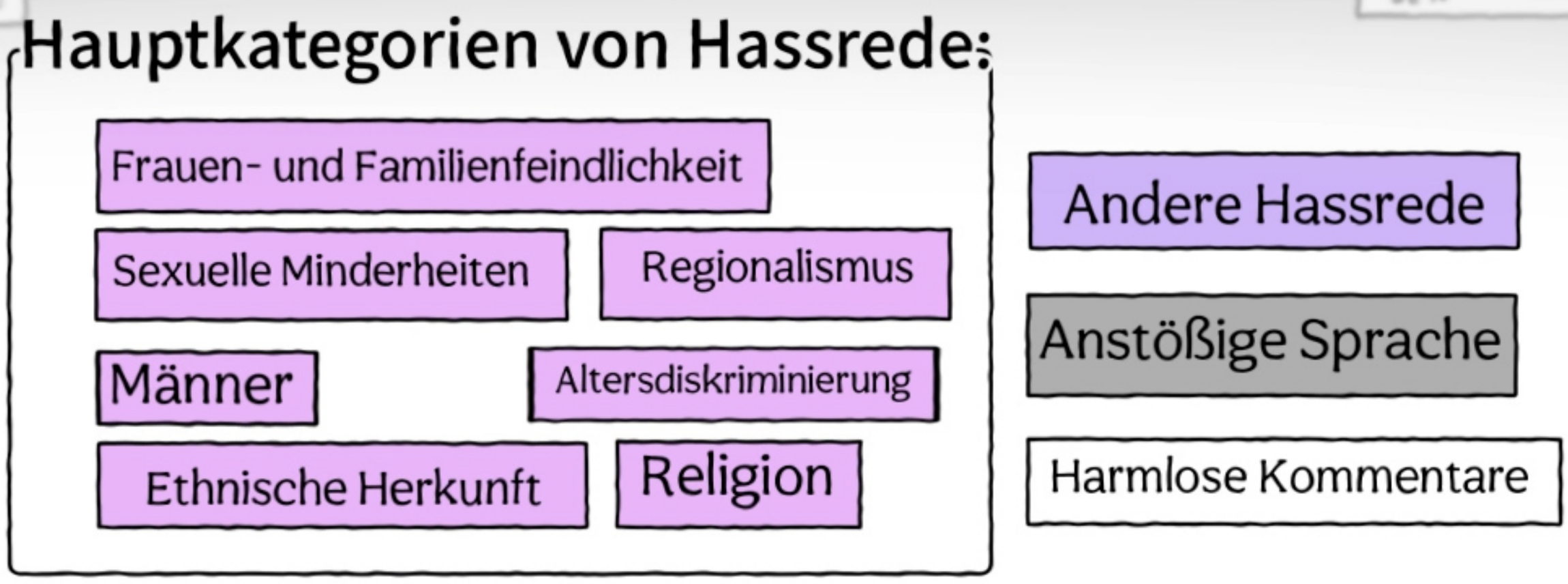
Tae Young Kang
Gründer von Underscore



Unser Unsmile-Datensatz enthält viele Textbeispiele, die KI-Trainingsmodellen dazu dienen können, Hassrede mit großer Genauigkeit zu erkennen.

Aber wie funktioniert es eigentlich?

Der Datensatz beruht auf über 23.000 Kommentaren aus wichtigen Online-Nachrichtenquellen und von Community-Sites. Diese wurden den folgenden 10 Gruppen zugeordnet.

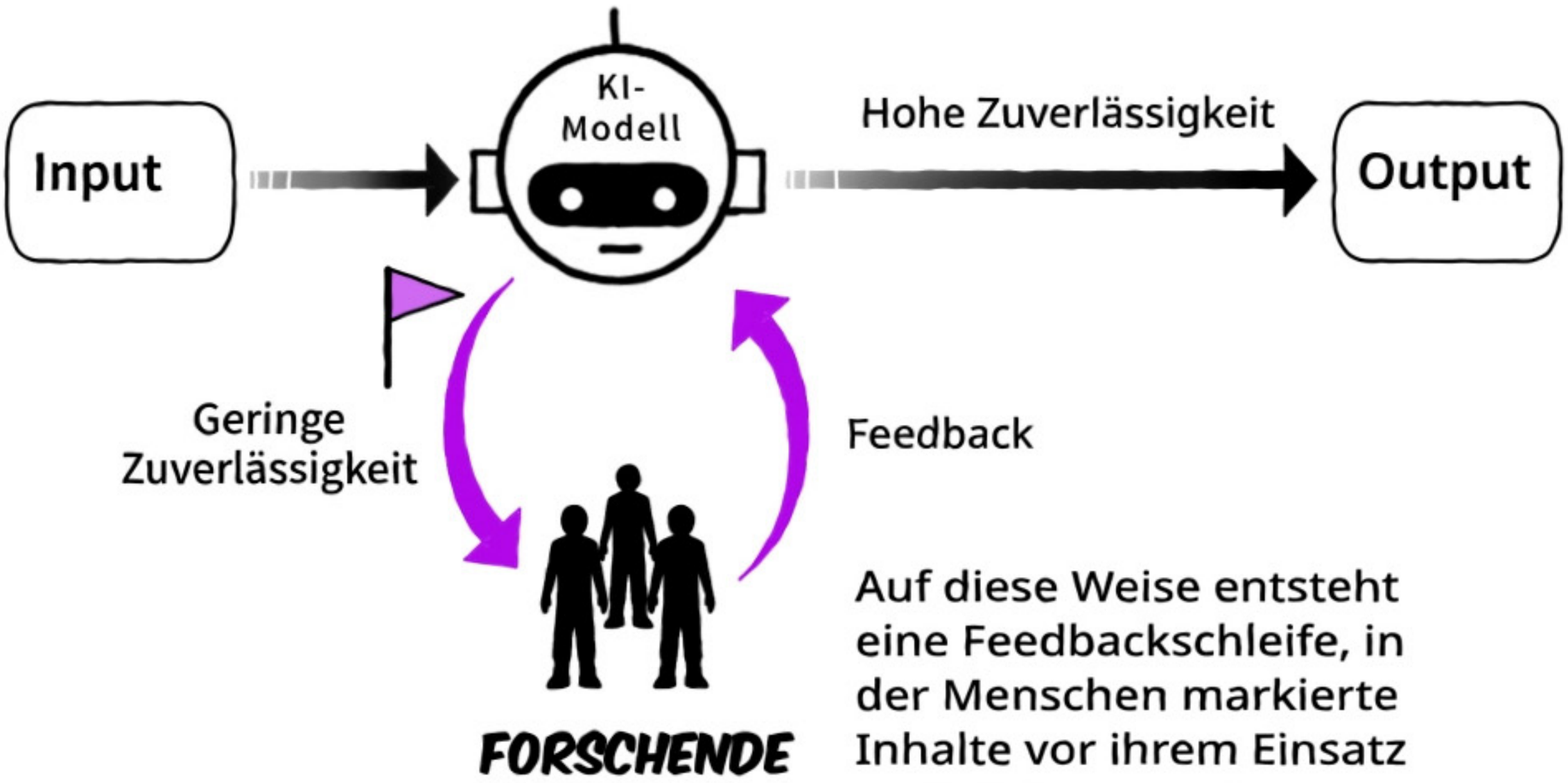


In schwierigen Fällen helfen Forschende mit der Einordnung der Kommentare.

Anschließend wird der Datensatz von einem parallel laufenden Algorithmus namens „Hatescore“ analysiert. Er soll berechnen, wie wahrscheinlich es ist, dass etwas als Hassrede wahrgenommen wird.

Die Idee besteht darin, Hatescore in KI-Trainingsmodelle zu integrieren, um Hassrede zu ermitteln und sofort zu blockieren.

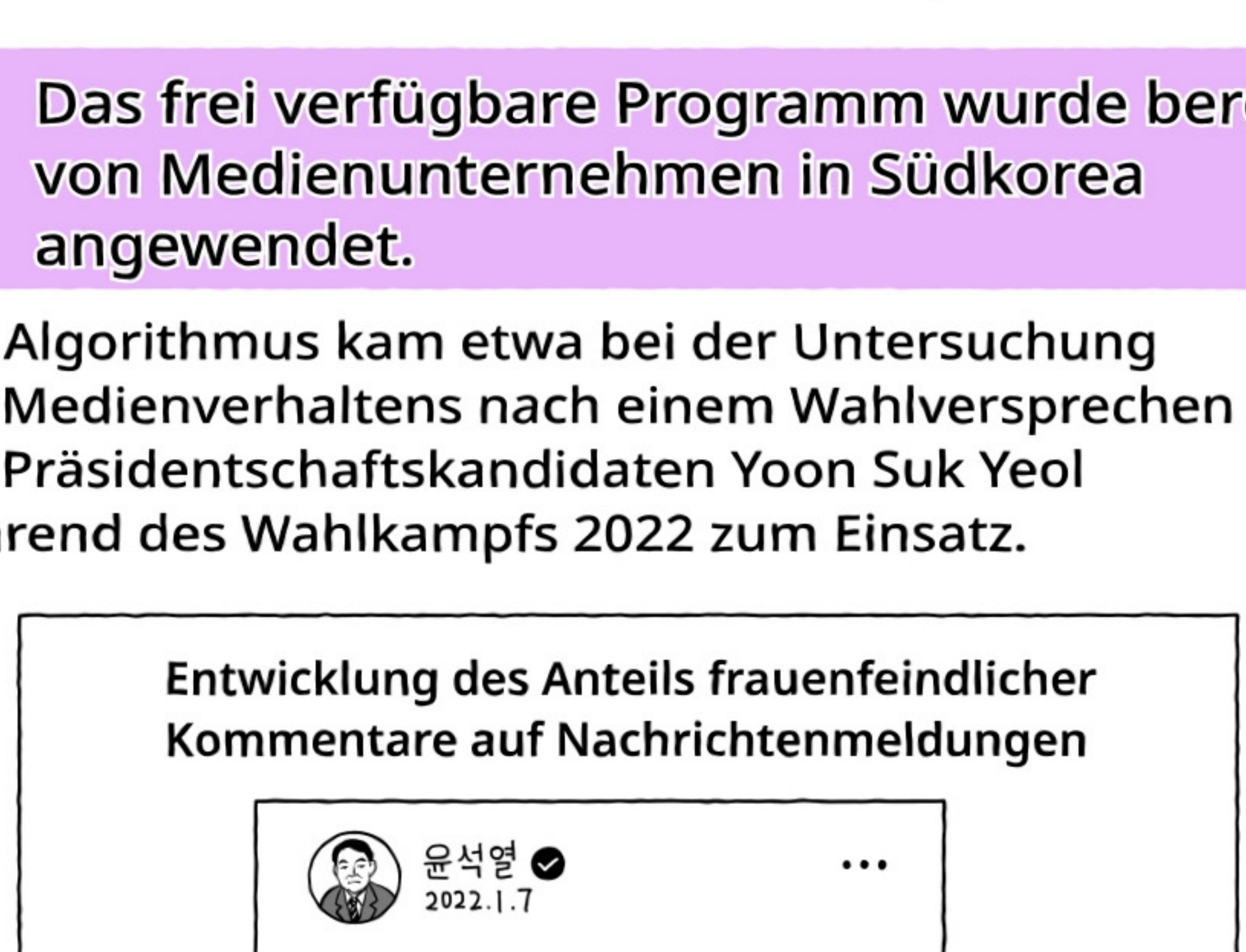
Der Algorithmus kann Kommentare mit unzuverlässigen Ergebnissen zur weiteren Prüfung markieren.



Auf diese Weise entsteht eine Feedbackschleife, in der Menschen markierte Inhalte vor ihrem Einsatz zu Trainingszwecken prüfen. Dadurch kann sich das System laufend an neue Muster von Hassrede anpassen.

Das frei verfügbare Programm wurde bereits von Medienunternehmen in Südkorea angewendet.

Der Algorithmus kam etwa bei der Untersuchung des Medienverhaltens nach einem Wahlversprechen des Präsidentschaftskandidaten Yoon Suk Yeol während des Wahlkampfs 2022 zum Einsatz.



In diesem Fall lieferte die Hatescore-Analyse Daten, die die Recherche der Zeitung stützten.

Allerdings weisen Forschende darauf hin, dass Hassrede nicht nur durch KI herausgefiltert werden sollte. Das könne wichtige Informationen einschränken und zu falschen Einordnungen führen.

Umfang des Unsmile-Datensatzes

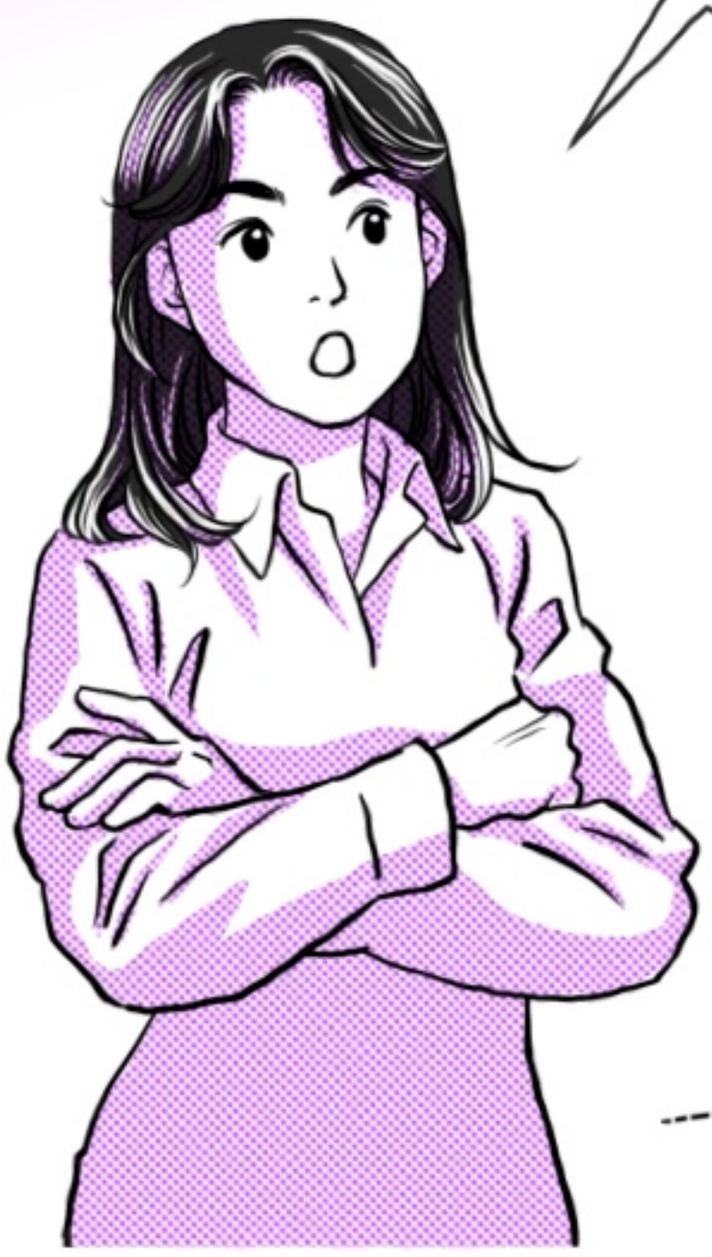
- Frauen- und Familienfeindlichkeit
- Sexuelle Minderheiten
- Männer
- Ethnische Herkunft & Nationalität
- Regionalismus
- Altersdiskriminierung
- Religion

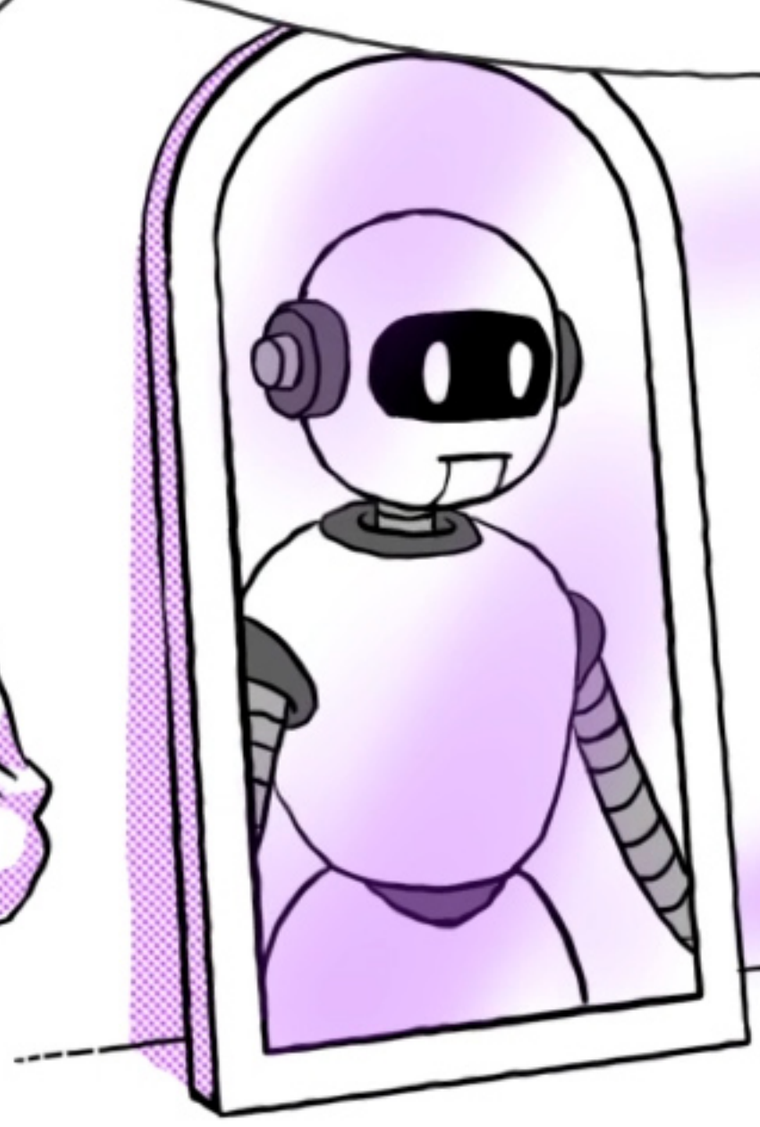
Other hate speech

KI-Systeme wie Luda spiegeln gesellschaftliche Werte und Vorurteile wider.

Das muss während der gesamten Planung, Gestaltung, Entwicklung und Nutzung von KI mitgedacht werden.

Yubeen Kwon
Forschungsassistentin an der Seoul National University





KI-Systeme halten immer mehr Einzug in den Alltag. Deshalb wird es zur gesellschaftlichen und politischen Herausforderung, dass sie ethische Standards einhalten – und zwar ohne die Meinungsfreiheit einzuschränken oder Vorurteile widerzuspiegeln.

Um faire, transparente und verantwortungsvolle Systeme zu schaffen, müssen verschiedenste Interessensgruppen aus Sozialwissenschaften, Ethik und der Gesellschaft gehört werden.

bpb.de Den Gerüchten auf der Spur

Search


bpb:
GOETHE
INSTITUT