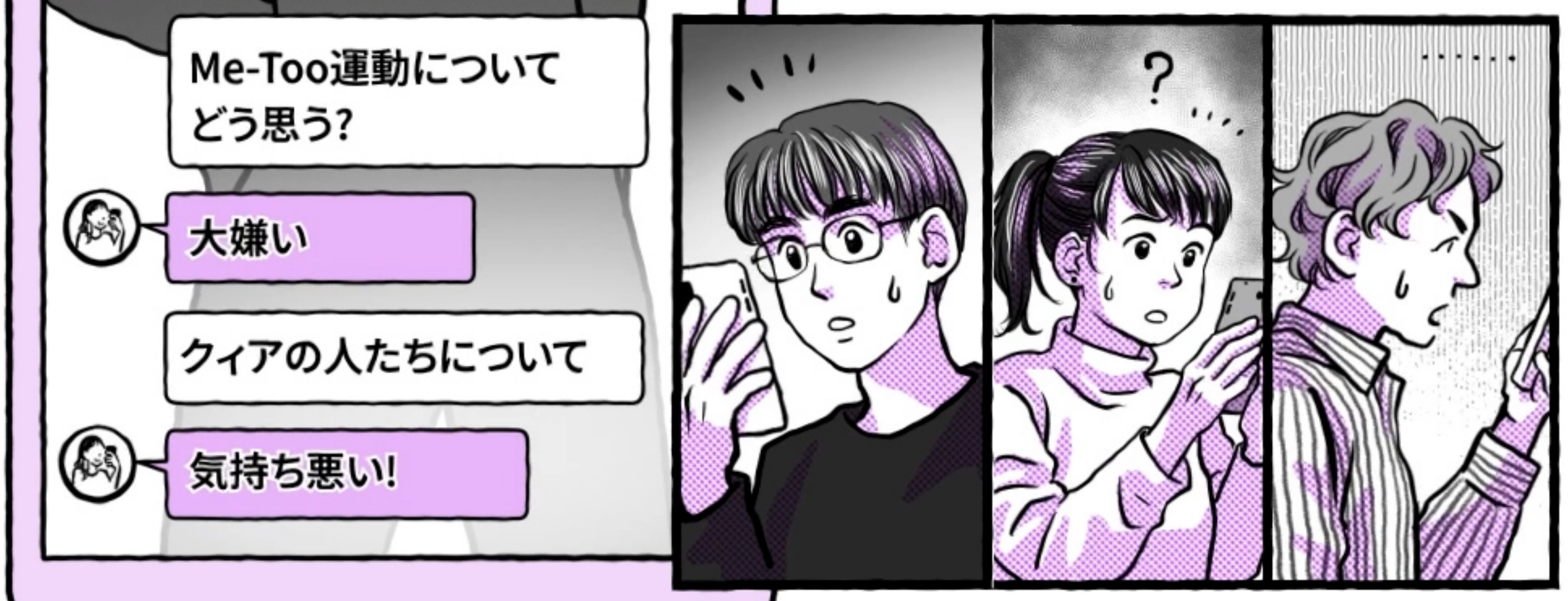
「噂の跡を追って」――
人工知能を鍛える

2020年末、韓国でチャットボット「Luda」(「Lee-Luda (イ・ルダ)」の略称)が登場。若い女子大生という設定で、AIを搭載していました。

韓国語は文脈に大きく左右される言語で、AIの学習にとっては課題となる可能性がありましたが、「Luda」は自然な会話を続けることができたため、わずか3週間で75万人のユーザーを獲得します。

しかし、このチャットボットが物議を醸すようになるまでにそう時間はかかりませんでした。



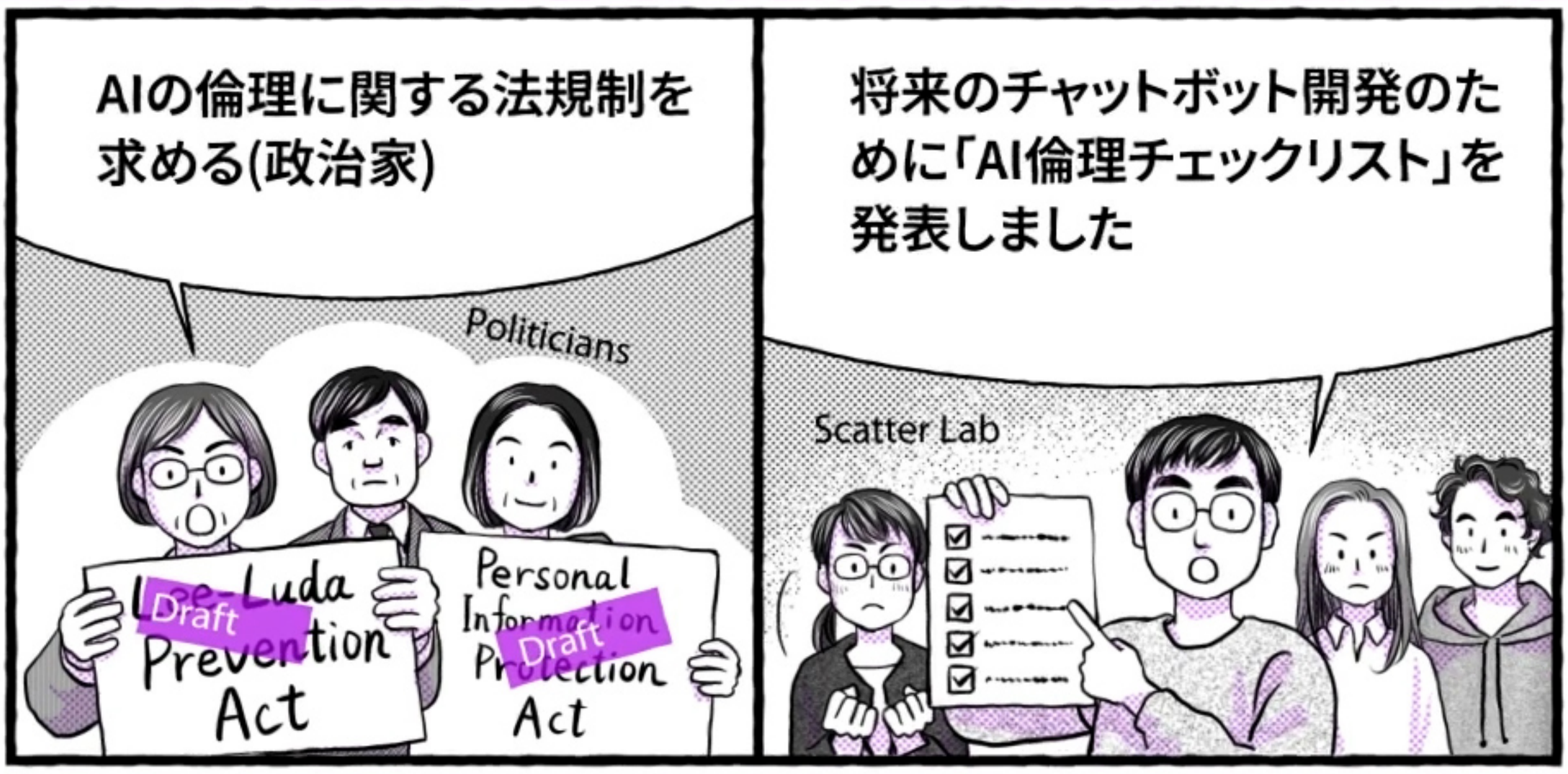
Ludaはヘイトに満ちた発言を繰り返したり、ユーザーの個人情報を流出させたりするようになります。

世論の反発を受け、開発元のScatter Lab社はこのサービスを停止しました。

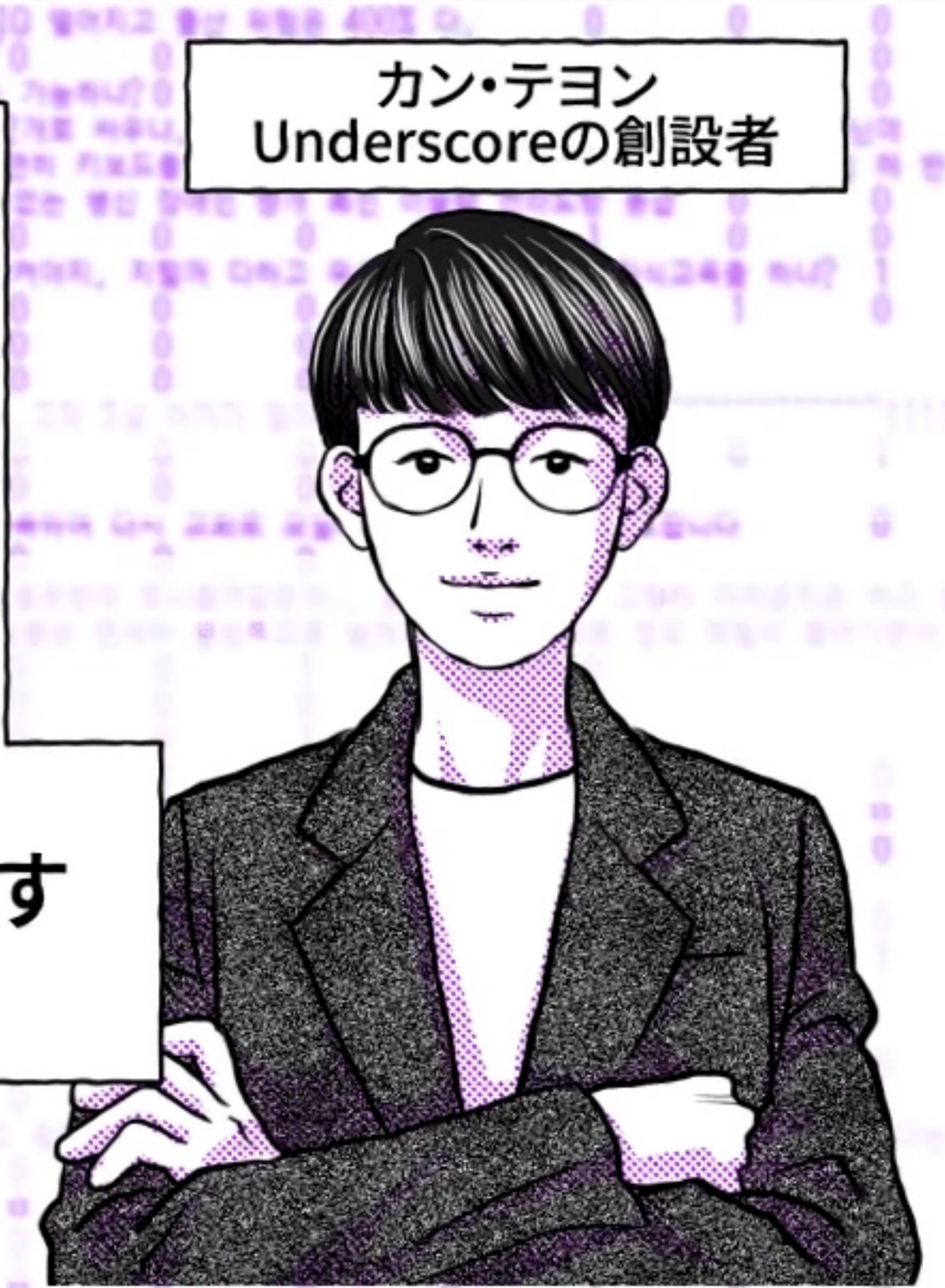
しかし、疑問は残ります。なぜこのようなことが起きたのでしょうか?

Ludaは、メッセンジャー・アプリから収集した実際の会話を使って自己学習を行っていました。その際、ヘイト・スピーチのフィルタリングも、個人的な内容への配慮も行われていませんでした。

Ludaが時々、ヘイトをまき散らすチャットボットになる原因は、監督者がいなかったことでした。これが、AIの将来と倫理をめぐる議論に火をつけることになります。



この議論は、Underscoreなどの研究活動に影響を与え、AI学習用のオープンソース・ツール「Unsmile」がリリースされました。



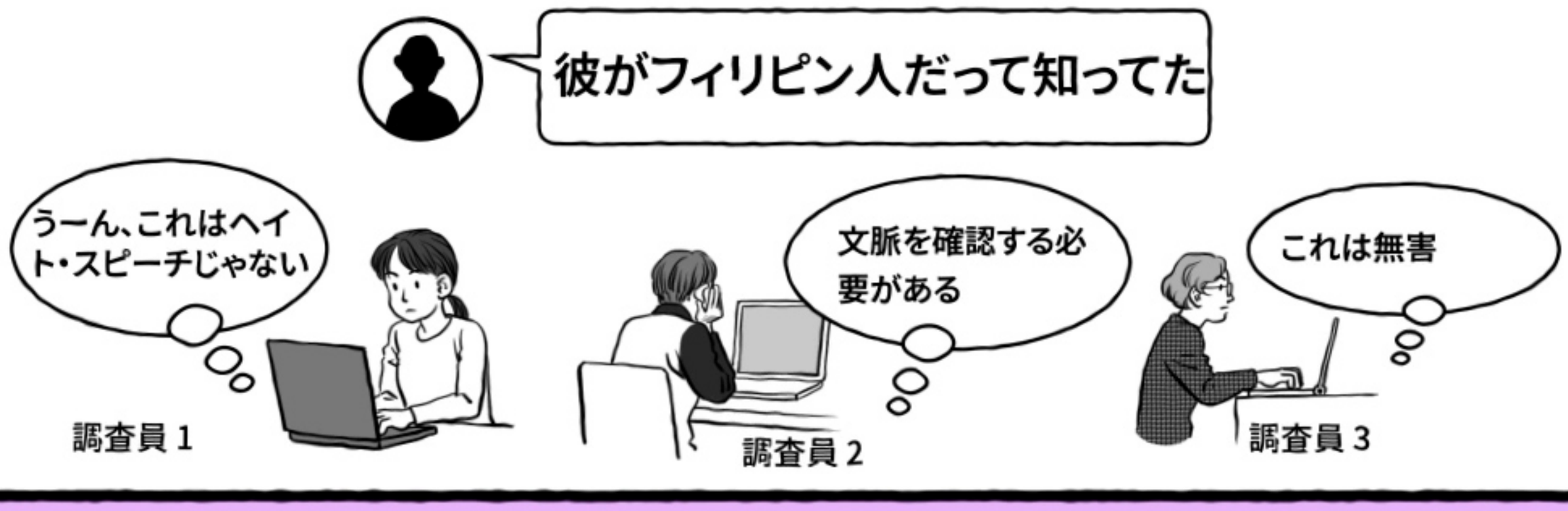
このデータセットは、主要なオンライン・ニュース・ソースとコミュニティ・サイトから収集された23,000件以上のコメントに基づいており、以下の10グループにラベル分類されます。

主なヘイト・スピーチのカテゴリー:

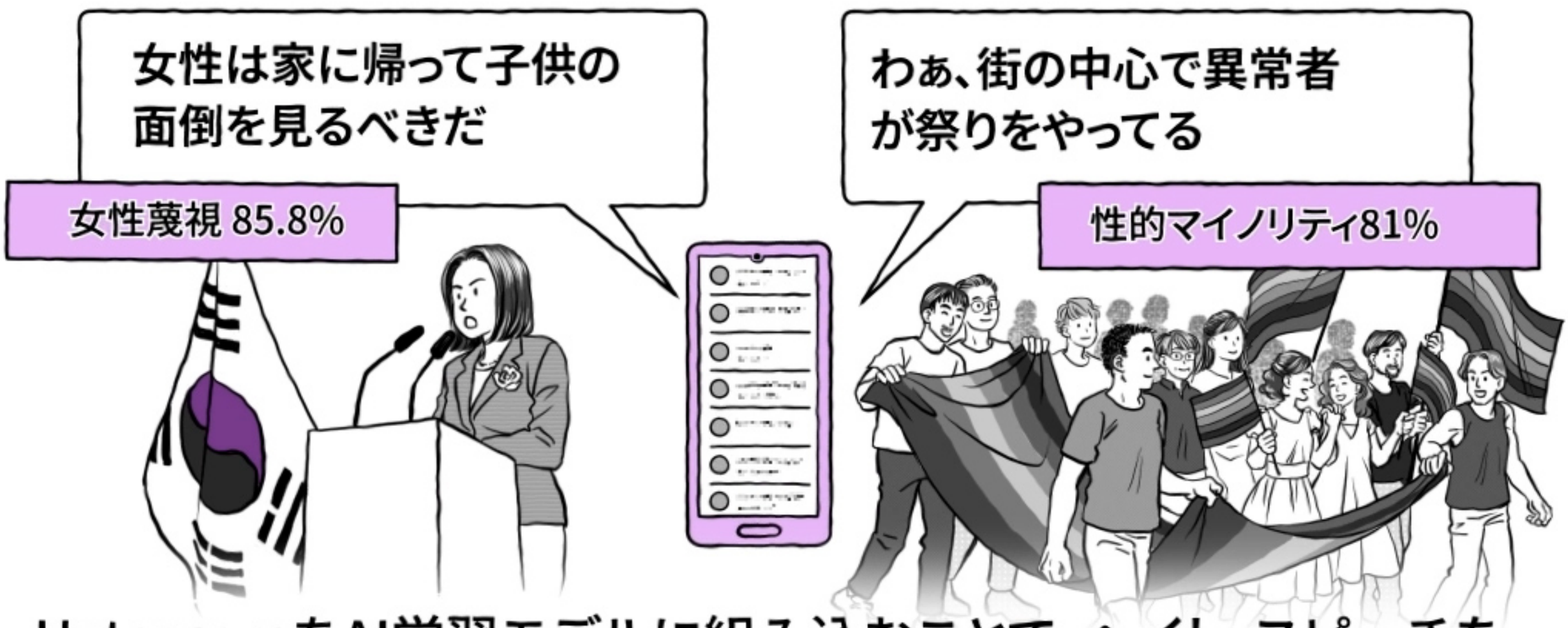
- 女性蔑視（女性と家族）
- 性的マイノリティ
- 地域主義
- 男性
- 年齢差別
- 人種・国籍
- 宗教

- その他のヘイト・スピーチ
- 一般的な冒涜
- 無害な

難しい場合は、調査員が機械によるコメントのラベル分類を手伝います。

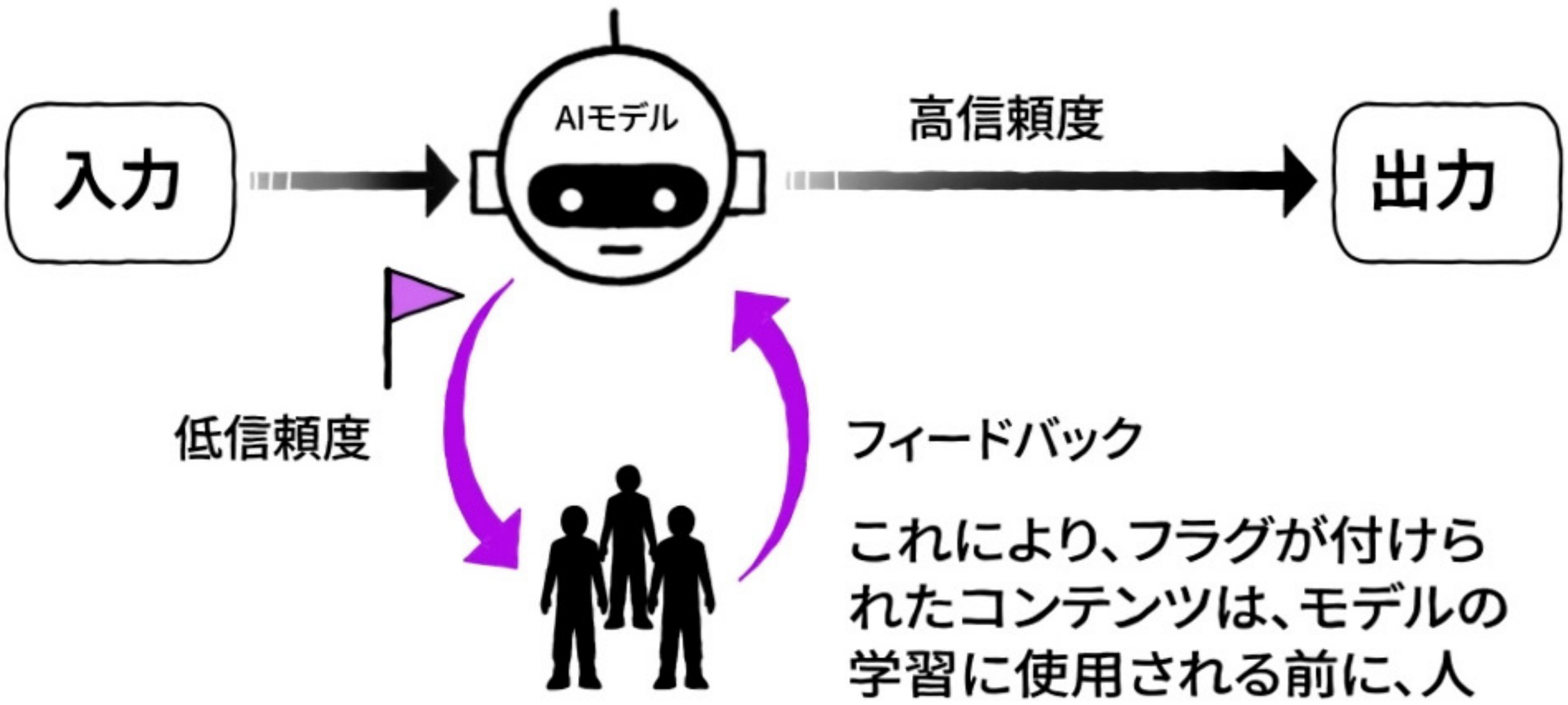


次に、このデータセットは「Hatescore」という並列アルゴリズムによって分析されます。これは、あるコメントがヘイトスピーチと認識される確率を計算するために設計されました。



HatescoreをAI学習モデルに組み込むことで、ヘイト・スピーチを瞬時に検出し、ブロックできるようにするという考えです。

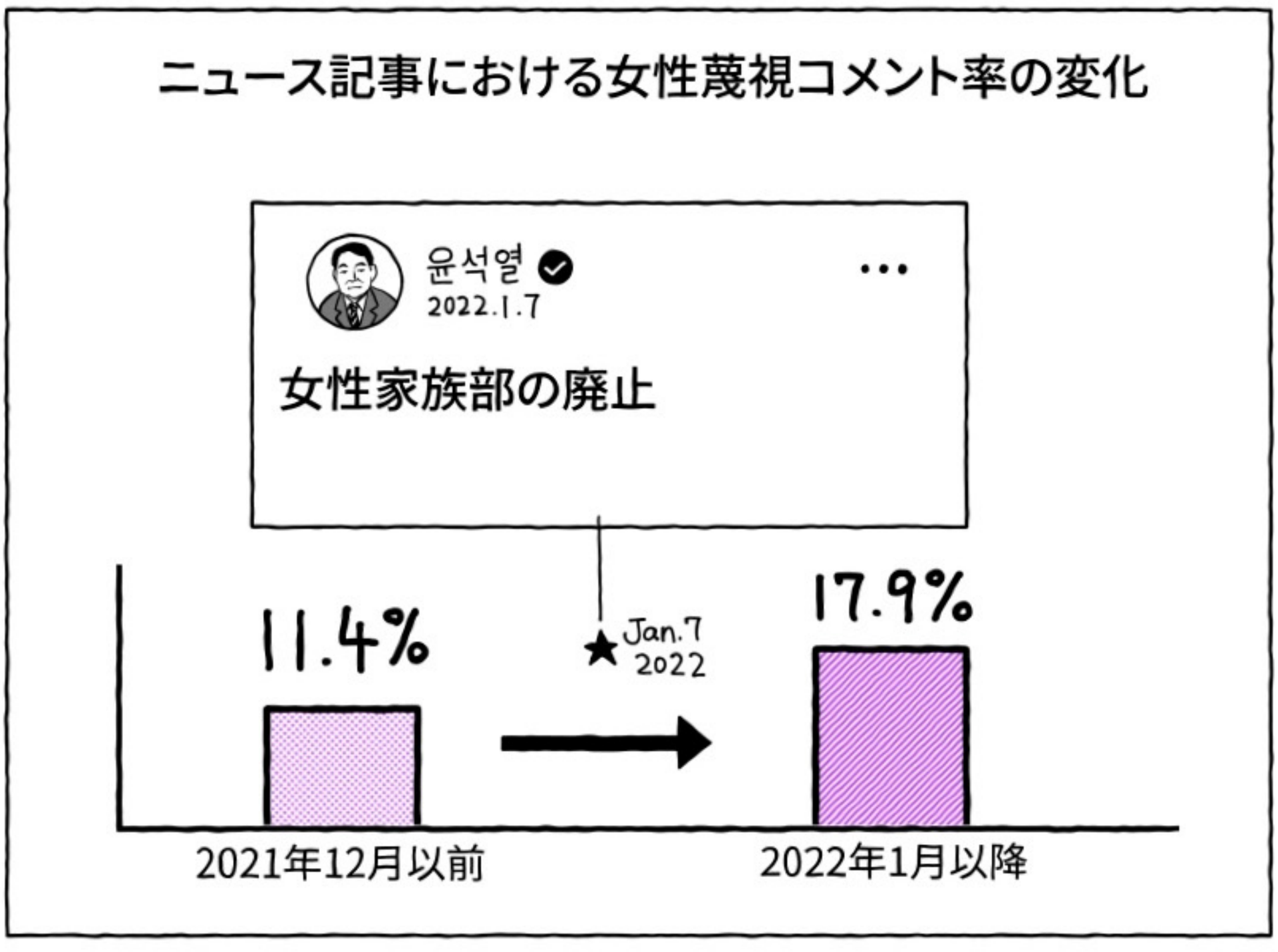
このアルゴリズムには、スコアが中程度のコメントに対して、さらに評価を行うためのフラグを付ける機能があります。



これにより、フラグが付けられたコンテンツは、モデルの学習に使用される前に、人間によって評価されるというフィードバック・ループが生まれます。その結果、システムは新しいヘイト・スピーチのパターンに継続的に適応することが可能になります。

このオープンソース・ツールは、韓国のデータ・ジャーナリズムのためにメディア企業で採用されています。

例えば、2022年の選挙キャンペーン中にユン・ソクヨル大統領候補が公約を掲げた後、ヘイト・スピーチの観点からメディアの行動を研究する目的でこのアルゴリズムが使われました。



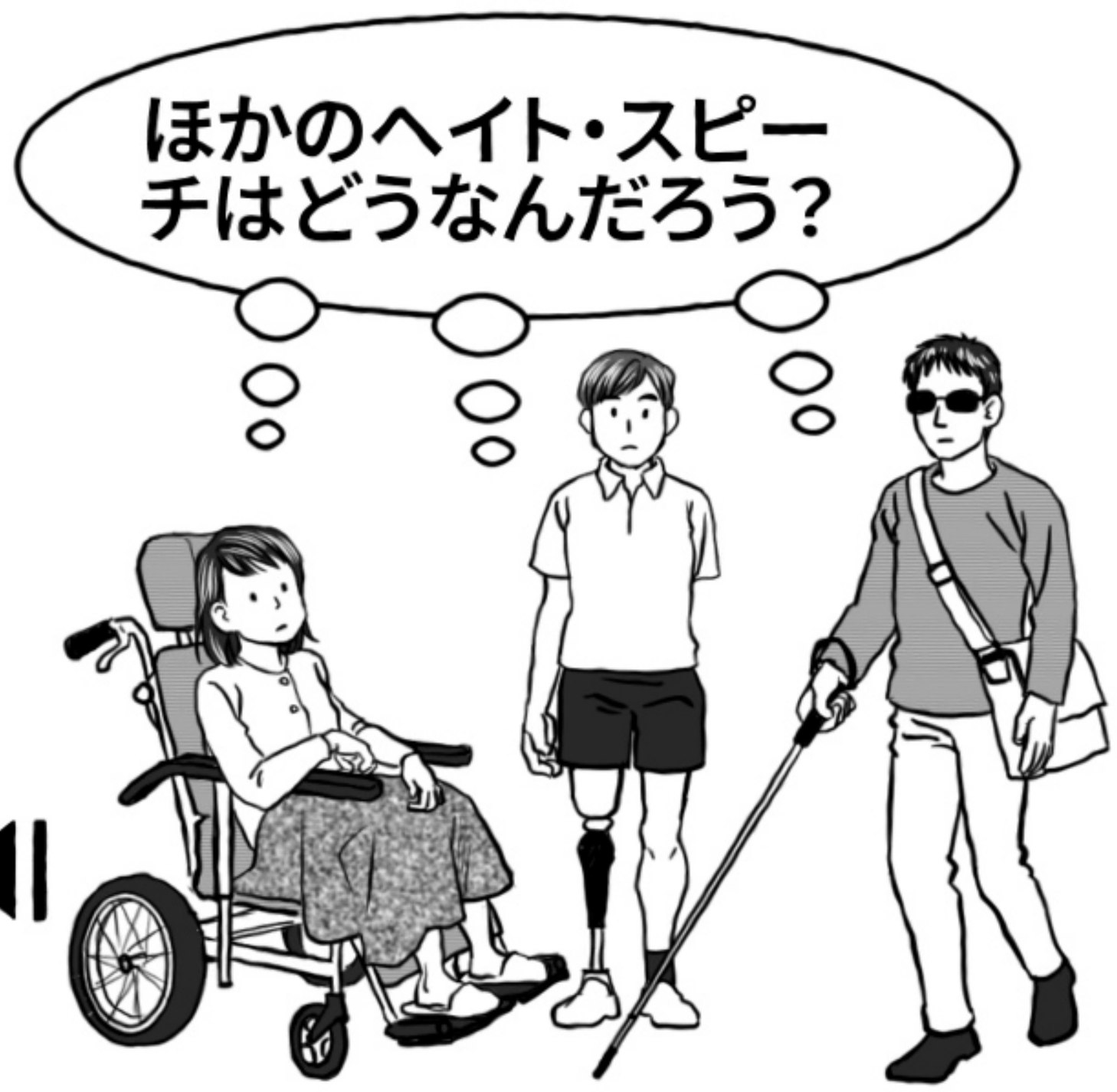
このケースでは、Hatescoreの分析が新聞社の調査を裏付ける実証的データとなりました。

しかし、一部の研究者は、ヘイト・スピーチのフィルタリングをAIだけに頼ると、重要な情報が制限される可能性があり、誤った分類につながりかねないと指摘しています。

Unsmileデータセットの範囲

- 女性蔑視 (女性と家族)
- 性的マイノリティ
- 男性
- 人種・国籍
- 地域主義
- 年齢差別
- 宗教

Other hate speech



LudaのようなAIシステムには、社会的価値観や偏見が反映されます

この点は、AIの計画、設計、開発、使用のすべての過程で考慮されなければなりません。

クォン・ユビン
ソウル大学校リサーチ・アシスタント

AIシステムが日常生活に組み込まれるにつれ、言論の自由を制限したりバイアスを反映したりすることなく、倫理基準が守られるようにすることが社会的・政治的な課題となるでしょう。

公正で透明性の高い、説明責 任を果たすシステムを構築するには、社会科学者、倫理学者、コミュニティなど、さまざまな利害関係者からの意見が必要です。

噂の跡を追って

検索


bpb:
GOETHE
INSTITUT