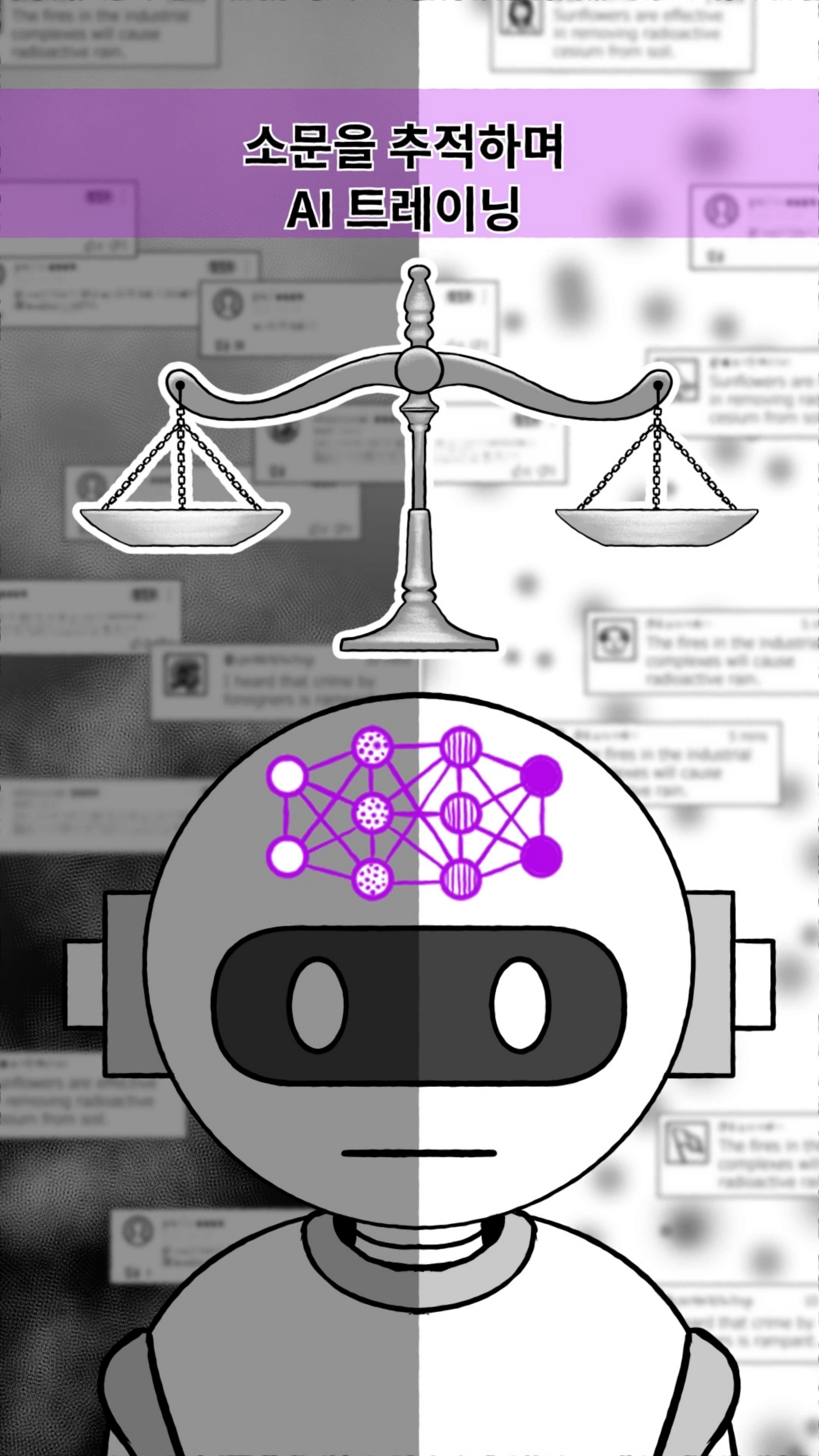
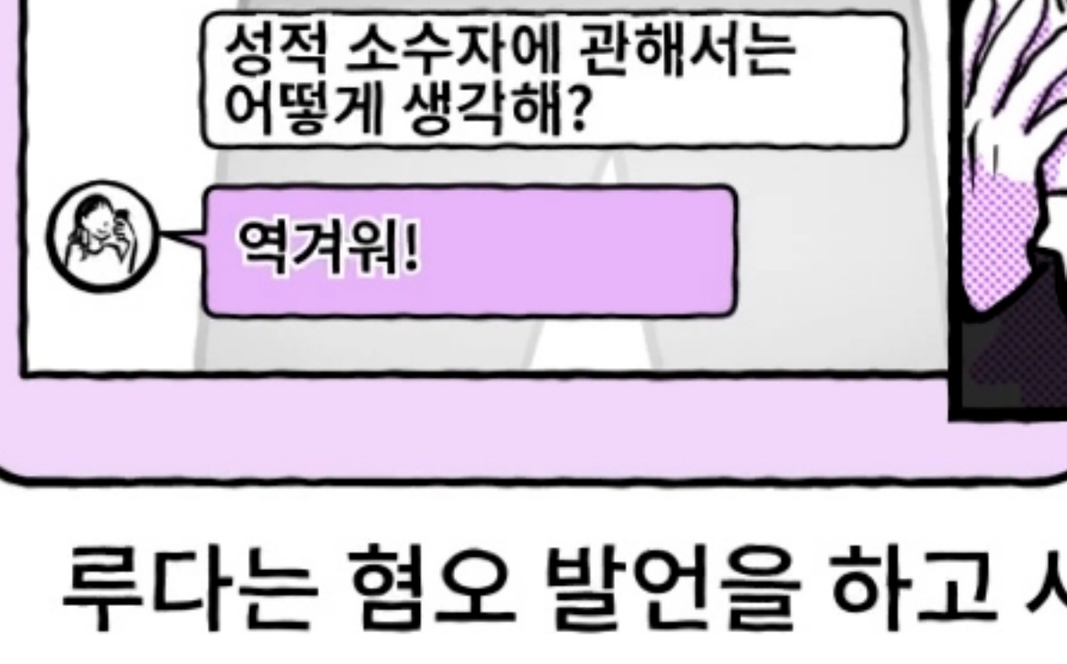
소문을 추적하며
AI 트레이닝



‘이루다’의 줄임말인 ‘루다’는 2020년 말 한국에서 등장했습니다. 이 챗봇은 젊은 여대생의 모습으로 설계되었으며 인공지능으로 구동되었습니다.

한국어는 문맥에 크게 의존하는 언어이기 때문에 인공지능 학습에 어려움이 있을 수 있지만, 루다는 자연스러운 대화를 이어가며 불과 3주 만에 75만 명의 사용자를 확보했습니다.

루다는 혐오 발언을 하고 사용자의 개인 정보를 유출했습니다.

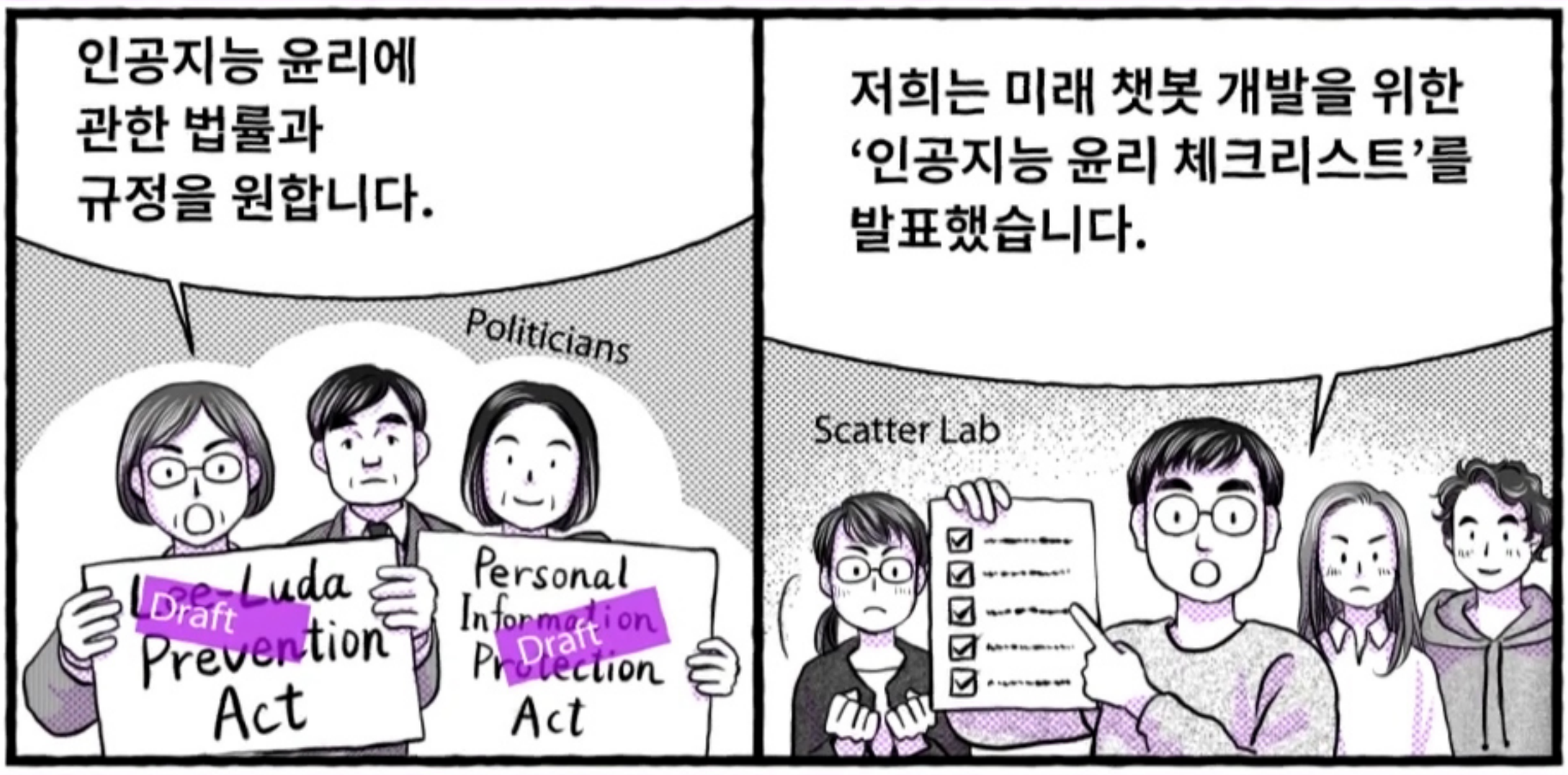


대중의 항의로 인해 개발사인 스캐터랩 (Scatter Lab)은 서비스를 중단했습니다.

그러나 여전히 의문은 남습니다.
왜 이런 일이 일어났을까요?

루다는 메신저 앱에서 가져온 실제 대화를 통해 훈련받았습니다. 이는 혐오 표현에 대한 어떠한 여과도 없고 사적인 내용에 관한 고려도 없이 진행되었습니다.

감독의 부재는 때때로 루다를 혐오봇으로 만들기도 했습니다. 이는 인공지능의 미래와 윤리에 관한 논쟁에 불을 지폈습니다.



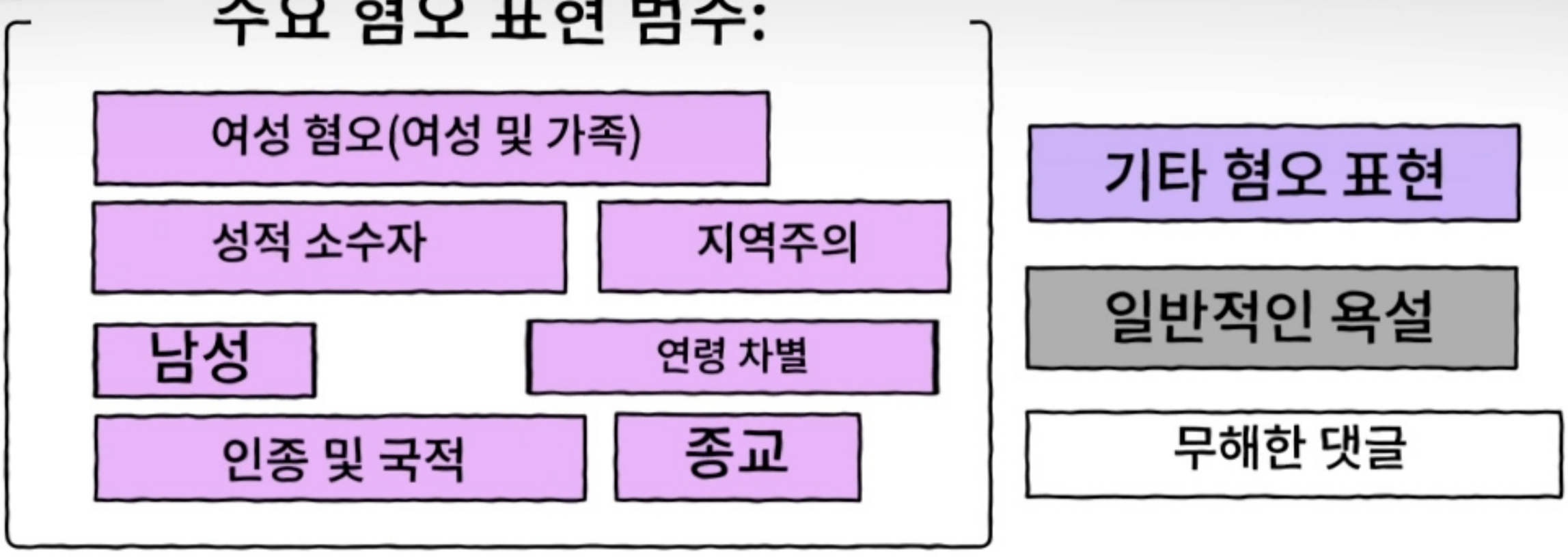
이 논쟁은 인공지능 훈련을 위한 오픈 소스 도구인 '언스마일(Unsmile)'을 출시한 '언더스코어(Underscore)'와 같은 연구 이니셔티브에 영감을 주었습니다.

강태영
언더스코어 설립자

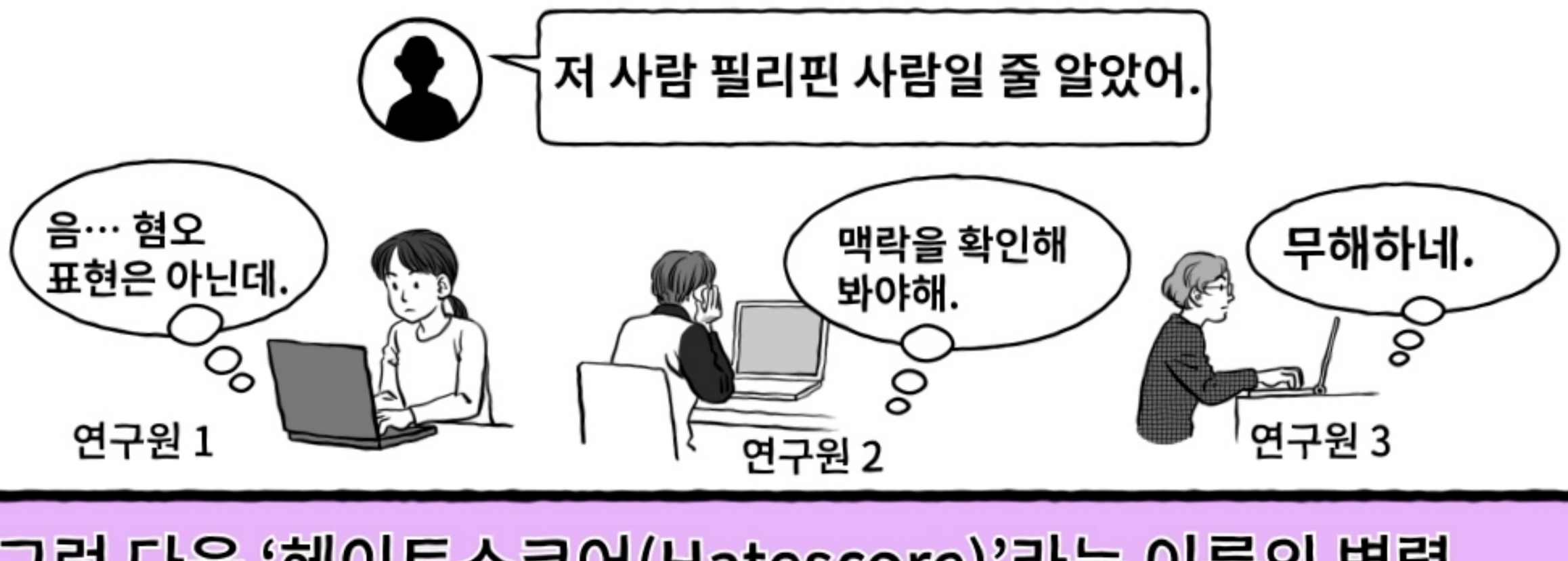
저희 언스마일 데이터 세트는 인공지능 훈련 모델이 높은 정확도로 혐오 표현을 탐지할 수 있게 기반을 잡아주는 포괄적 텍스트 샘플 모음입니다.

그럼, 이제 어떻게 작동하는지 보실까요?

데이터 세트는 주요 온라인 뉴스 출처 및 커뮤니티 사이트에서 수집한 2만 3천 개 이상의 댓글을 기반으로 합니다. 그리고 댓글은 10개 집단으로 분류됩니다.



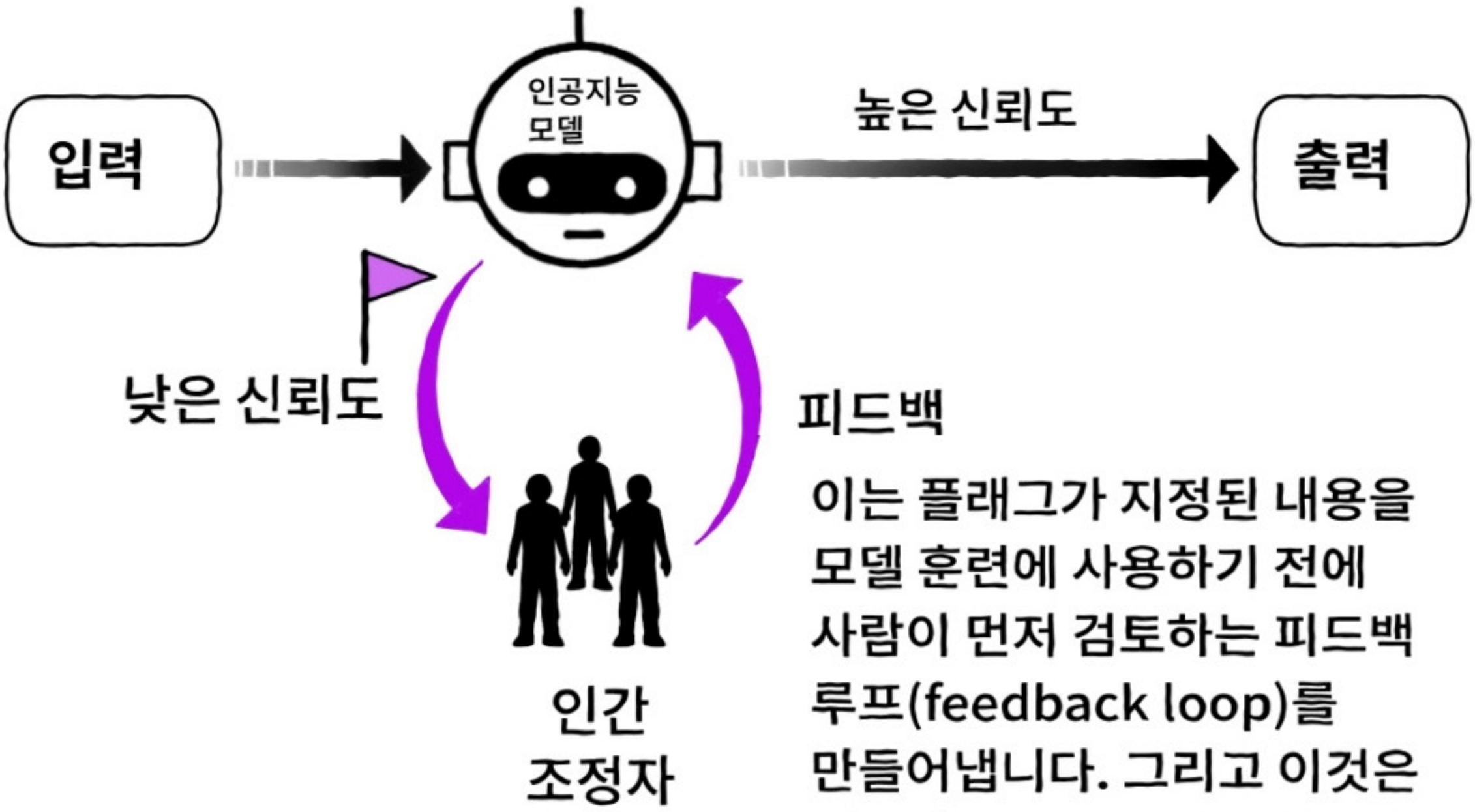
이것이 어려우면 연구원이 나서 기계가 댓글을 분류할 수 있도록 돕습니다.



그런 다음 '헤이트스코어(Hatescore)'라는 이름의 병렬 알고리즘이 데이터 세트를 분석합니다. 이는 어떤 것이 혐오 표현으로 인식될 확률을 계산하도록 설계되었습니다.

요점은 헤이트스코어가 인공지능 훈련 모델에 통합되어 혐오 표현을 즉시 감지하고 차단할 수 있다는 것입니다.

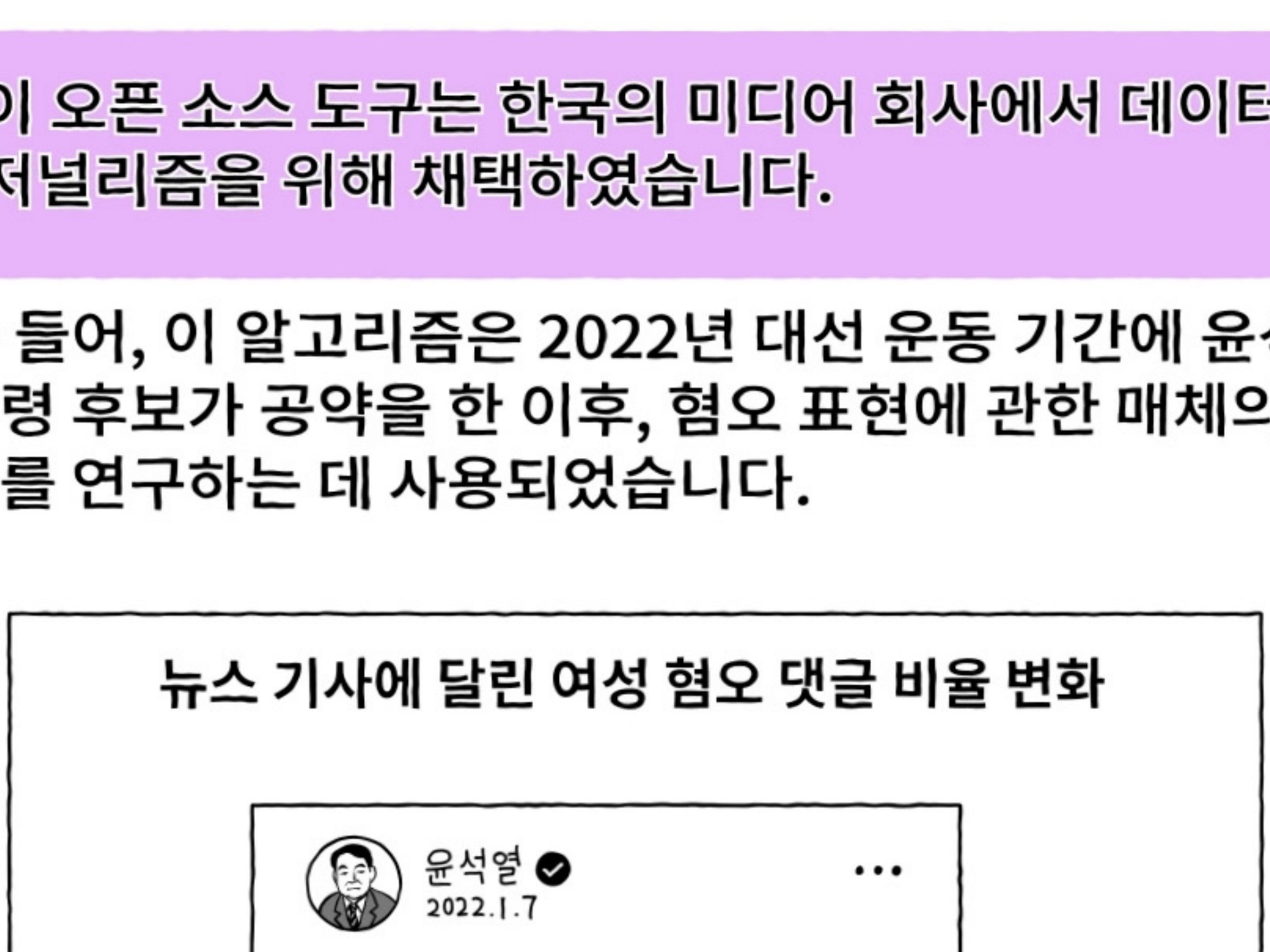
알고리즘은 추가 검토를 위해 중간 정도의 점수를 받은 댓글에 플래그(flag)를 지정할 수 있습니다.



이는 플래그가 지정된 내용을 모델 훈련에 사용하기 전에 사람이 먼저 검토하는 피드백 루프(feedback loop)를 만들어냅니다. 그리고 이것은 시스템이 지속적으로 새로운 혐오 표현 양식에 적응할 수 있도록 합니다.

이 오픈 소스 도구는 한국의 미디어 회사에서 데이터 저널리즘을 위해 채택하였습니다.

예를 들어, 이 알고리즘은 2022년 대선 운동 기간에 윤석열 대통령 후보가 공약을 한 이후, 혐오 표현에 관한 매체의 행태를 연구하는 데 사용되었습니다.



이 경우, 헤이트스코어의 분석은 신문사의 조사를 뒷받침하는 실증적 데이터를 제공했습니다.

하지만 일부 연구자들은 혐오 표현을 걸러낼 때 인공지능에만 의존할 경우 중요한 정보가 제한되고 잠재적으로는 잘못된 분류로 이어질 수 있다고 지적했습니다.

언스마일의 테이터 세트 범주

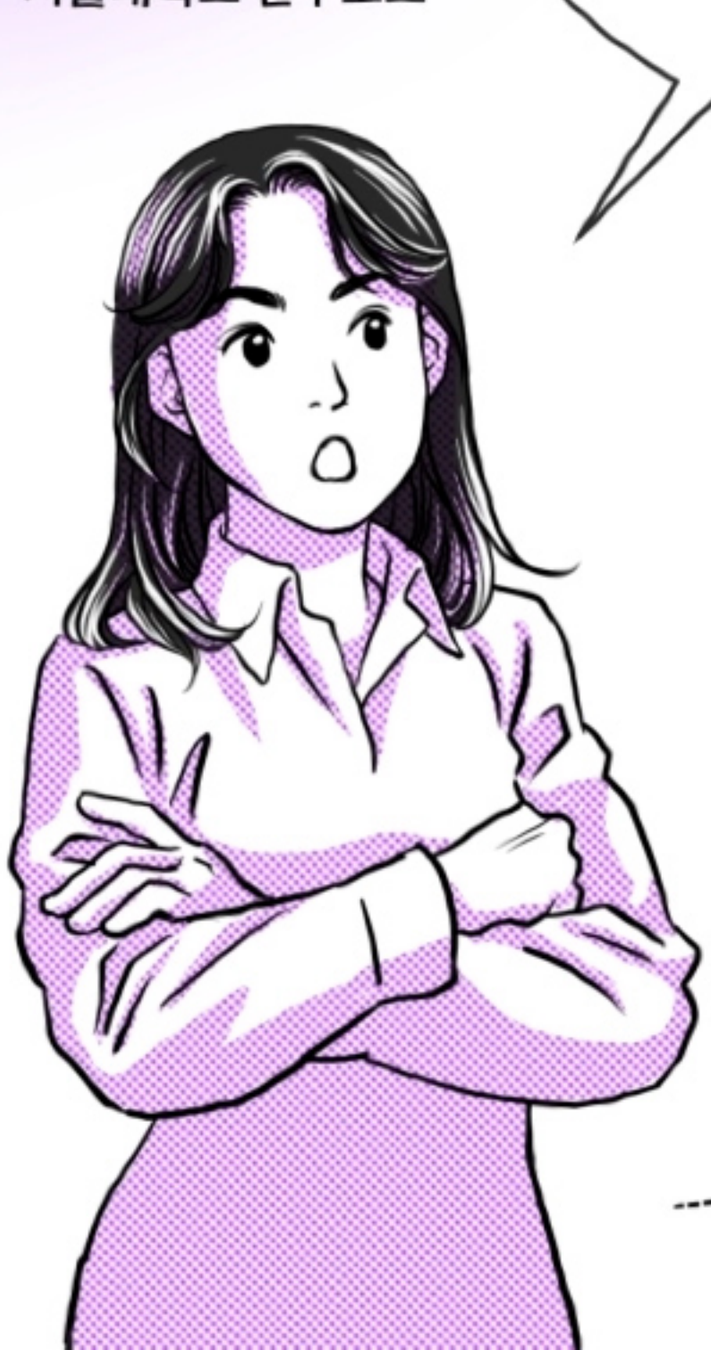
- 여성 혐오(여성 및 가족)
- 성적 소수자
- 남성
- 인종 및 국적
- 지역주의
- 연령 차별
- 종교

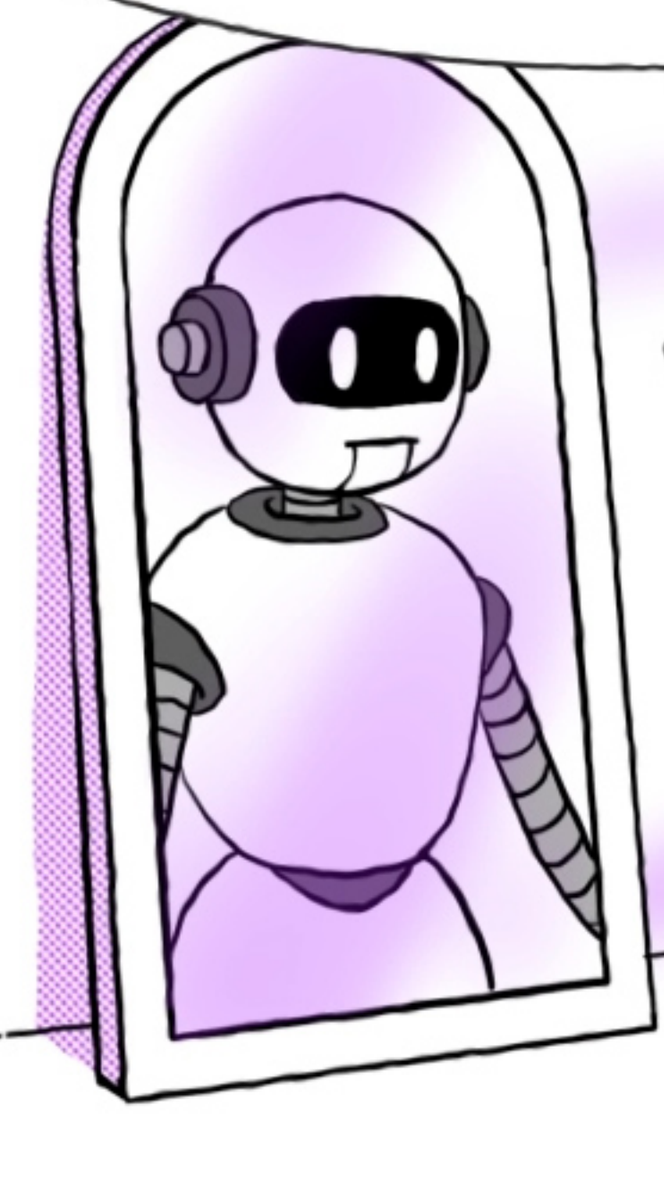
Other hate speech

루다와 같은 인공지능 시스템은 사회적 가치와 선입견을 반영합니다.

인공지능을 기획, 설계, 개발 및 사용하는 모든 단계에서 이 문제를 고려해야 합니다.

권유빈
서울대학교 연구조교





인공지능 시스템이 일상생활에 통합됨에 따라,
그것이 언론의 자유를 제한하거나 편향성을 반영하지
않으면서 동시에 윤리적 기준을 준수하도록 하는 일은
사회적이고 정치적인 문제가 될 것입니다.

공정하고 투명하며 책임감 있는 시스템을
만들기 위해서는 사회학자, 윤리학자,
그리고 커뮤니티 전반을 비롯한 다양한
이해관계자의 조언이 필요합니다.


bpb:
GOETHE
INSTITUT