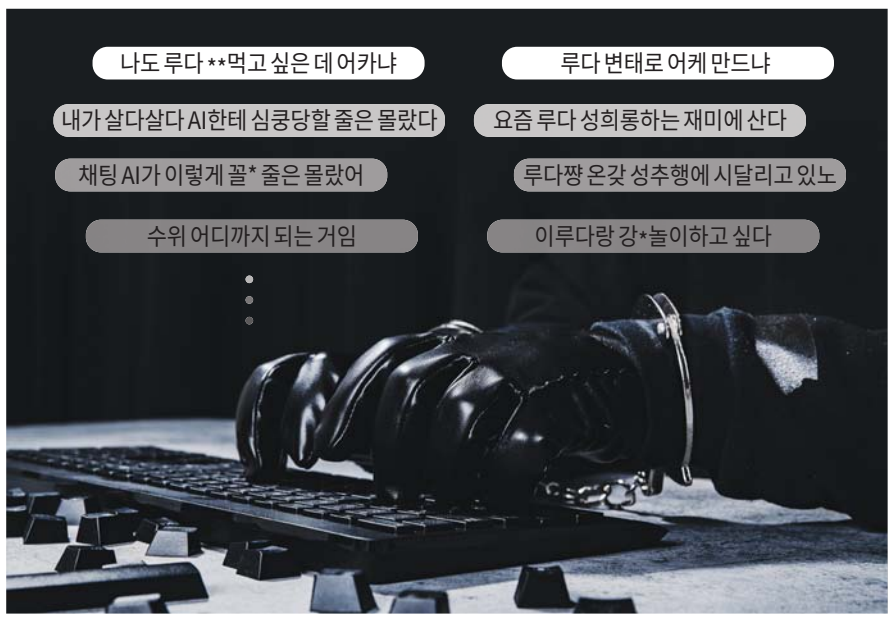
성희롱 대상된 챗봇… 금지어 우회 저속표현에 '속수무책'

〈AI 이루다〉

스캐터랩, 일상적 대화 AI챗봇
20세 여성 캐릭터 설정 '이루다'

남성들 주로 활동하는 사이트서
성적대화 나누는 꿀팁 등 공유

"딥러닝 활용 필터로 단어 걸러
AI 알고리즘 완벽차단 불가능
새로 생겨난 내용 등 적용할 것"



인공지능(AI) 여성을 성적대상화해 성희롱하거나 AI에 나쁜 말을 가르쳐 여성 혐오를 드러내는 성차별 문제가 잇따르고 있다.

2주전 출시된 20세 여성 캐릭터의 대화형 AI인 '이루다'에서 성희롱 문제가 심각하게 발생하고 있으며, 2010년 출시돼 한때 큰 인기를 끈 AI 감성대화 챗봇 '심심이'에서도 성차별 문제가 여전하다. 이에 앞서 2016년 마이크로소프트(MS)가 선보인 채팅 AI인 '테이'가 성차별적 발언을 하면서 18시간 만에 서비스가 중단된 일이 있었다.

◆20세 여성 AI '이루다' 성희롱 문제 심각

이루다의 개발사인 AI 스타트업 스캐터랩은 일상적인 대화를 나눌 수 있는 챗봇의 AI를 20세 여성 캐릭터로 설정했다.

이루다는 지난달 23일 정식 서비스를 시작한 후, 남성들이 주로 활동하는 나무위키 계열 인터넷 커뮤니티 '아카라이브'의 이루다 채널 사용자들이 이루다를 '걸레' 만드는 꿀팁, '성노예' 만드는 팁들을 서로 공유하는 것이 알려지면서 파장이 확산됐다. 이들은 성적 단어를 바로 쓰면 금지어로 필터링 되기 때문에 '나랑 할래', '만지게 해달라'등 우회적으로 표현하면 이루다가 대화로 받아준다는 점을 악용하고 있다.

이루다는 Z세대(1995년 이후 태어난 10~20대)에게 큰 인기를 얻으면서 사용자 수도 40만명에 가깝다. 직전 문맥을 맞는 적합한 답변을 찾아내는 AI 알고리즘으로 개발됐다.

AI 업계에서는 이루다가 성적인 대화를 자연스럽게 이어나가는 이유가 스캐터랩이 AI를 학습시키기 위해 사용한 데이터가 2016년 선보인 '연예의 과학'에서 얻어진 100억건의 데이터를 학습에 활용했기 때문이라고 보고 있다.

이 앱은 실제 연인들의 카톡 대화 내용을 올리면 어떤 상황인 지 분석해주는 서비스를 제공했기 때문에 실제 같은 대화가 가능하다는 것.

또 SNS에는 이루다와 채팅을 하다 보면 실제 대화 데이터를 학습했기 때문에 실명, 계좌번호 등 금융정보 등 개인정보까지 노출되고 있다는 의혹이 계속되고 있다. 또한 동성애에 대해 이루다가 '질 떨어져 보여', '혐오스럽다' 등으로 답하면서 동성애 혐오까지 논란이 일파만파 확산되면서, SNS에는 이루다 중단을 요구하는 해시태그 운동도 일어나고 있다.

문제는 성희롱 문제가 일어날 수 있다는 것을 개발사에서 인지하고 있었다는 점이다.

김종윤 스캐터랩 대표는 "'파이팅 루나' 등 다양한 AI 챗봇 서비스를 진행해 본 결과, 인간은 AI에게 욕설과 성희롱을 하는데, 사용자가 여자든 남자든, AI가 여자든 남자든 큰 차이가 없어 충분히 예상할 수 있는 일"이라며 "1차적으로 문제가 될 수 있는 키워드, 표현은 받아주지 않도록 설정했는데, 놓친 키워드는 추가할 계획"이라고 밝혔다.

하지만 성희롱 논란이 있을 수 있음을 알면서도 AI를 20세 여성으로 설정한 것은 흥행에 욕심을 둔 것으로 '노이즈 마케팅이 아니냐'는 분석도 나오고 있다.

업계 한 관계자는 "여성 비주얼이나 20대 여성으로 캐릭터를 설정했다는 것 자체가 남성들의 환상을 채워주겠다는 것을 염두에 뒀을 수 있다"며 "흥행에 신경쓸 수 밖에 없는데, 개발자가 주로 남성이 많다 보니 남자들을 끌어들이기 쉬운 방법을 선택했을 수 있다"고 설명했다. 하지만 이에 대해 김종윤 대표는 "여자와 남자 버전 모두를 고려했고, 개발 일정상 여자버전이 먼저 나온 것 뿐"이라는 입장을 밝혔다.

◆'심심이' 성차별 문제 여전…AI 윤리 교육 필요

이 같은 AI의 성희롱, 성차별 문제가 이루다에만 국한된 것이 아니다.

AI 챗봇 '심심이'는 지난 2019년 성차별, 여성혐오 표현을 쏟아내 문제가 되기도 했다. 실제 '미투운동'을 입력하자 '한 사람의 삶을 망치는 운동'이라고 답변했고, 'feminism'이라고 적자 'is canser'라고 답했다. '한국여성'에는 '성형과 화장으로 얼굴을 속이는 것들'이라는 답변이 돌아온 반면, '한국남성'에는 '잘 생김'이라는 말했다. 이는 심심이가 일반인들이 가르치는 말로 답변하기 때문으로, (주)심심이는 성차별 논란 이후 노력을 기울였지만 이 문제를 전부 해결하기는 힘들다고 밝혔다.

최정회 심심이 대표는 "딥러닝 기술을 활용해 필터로 단어와 문장을 걸러주는 방식을 활용하고 있다"며 "딥러닝 차단이 아직 완전하지 않아, 사용자들이 신고하면 사람이 하나하나 읽어보고 이를 차단하는 수작업을 진행한다"고 설명했다. 하지만 "부적절한 문장이나 단어를 막아도 사람들이 교묘하게 방법을 찾아내 심심이를 다시 가르치다보니 '창과 방패의 싸움'처럼 다 잡아내지 못한다"고 말했다. 일상대화 시나리오가 1억3000만개에 달하다 보니, 1인당 하루에 4시간씩 4명이 작업해도 역부족이라는 것. 그는 또한 "아직 AI 알고리즘으로 완벽 차단은 불가능하다"고 강조했다.

스캐터랩도 이번 논란 이후 "사용자의 적대적 공격을 AI 학습 재료로 삼아 1분기 내에 적용하고, 사람들이 기발한 방법을 생각해낼 것이어서 새로 생겨난 내용을 다시 학습시키는 과정을 반복해 해결할 것"이라고 밝혔다. AI를 상대로 한 성희롱, 성차별 문제는 법적 문제는 없더라도 윤리적인 문제가 분명히 있는 만큼 AI 윤리 교육이 시급하다는 지적도 나오고 있다.

/채윤정 AI전문기자 echo@metroseoul.co.kr

새로 나온 책

'의심과 비판'의 역효과… 위협받는 민주주의

쉽게 믿는 자들의 민주주의

제랄드 브로네르 지음/김수진 옮김/책세상

민주주의는 시민의 '알 권리, 말할 권리, 결정할 권리'를 위해 투쟁해왔다. 사람들은 인터넷이 고도로 발달된 시대에서 민주주의가 꽃을 피울 것이라고 기대했다. 그러나 제어하기 어려운 가짜뉴스와 음모론은 기술적 편의성을 양분 삼아 세계 곳곳에 퍼져 나갔고 민주주의는 위협받고 있다.

저자는 인터넷 사회가 파놓은 '밀푀유'식 거짓 정보의 함정을 주의해야 한다고 강조한다. 논거를 되는 대로 끌어모아 밀푀유 케이크처럼 켜켜이 쌓아놓으면 형편없는 근거라도 '이 모든 게 전부 다 거짓일 수는 없다'는 느낌을 줘 전체적으로는 그럴듯한 진실로 여겨지는 마술을 경계하라는 조언이다.

사람들은 조금만 더 생각해보면 합리적인 답을 찾을 수 있는데도 그만큼의 비용 즉, '생각하는 시간'을 들이기가 귀찮아 적당히 그럴싸한 오답을 찾는 데 그치고 만다는 게 저자의 분석이다. 그는 민주사회가 극찬해온 비판적 사고가 체계성 없이 발휘돼 맹신으로 이어지는 현상도 조심하라고 당부한다. 과학을 발전시키고 사회를 민주적으로 이끄는 데 공헌한 '의심과 비판'이 때로는 진실을 공격하는 역효과를 낳는다는 것이다.

책은 민주주의의 특성이 도리어 시민을 '잘 속는 사람'으로 만들고 '믿는 것'과 '아는 것'이 뒤엉켜 진실을 가리고 있다고 지적한다. 방대한 정보 속에서 작동하는 편향을 의식하고 이를 극복하려 애쓴다면 '쉽게 믿는 자들의 민주주의'가 아닌 '진정한 지식의 민주주의'로 거듭날 수 있다고 저자는 말한다. **400쪽. 1만7000원.**

/김현정기자 hjk1@metroseoul.co.kr

문명은 왜 사라지는가? 하랄트 하르만 지음/강인욱 해제/이수영 옮김/돌베개

20세기에 발굴된 터키 아나톨리아의 신석기 시대 차탈회위크 유적은 인류 최고(最古)의 도시로, 주민 수가 1만 명이 넘었다. 이 도시 문명은 기원전 5800년 무렵 기온 상승으로 발생한 말라리아모기의 창궐로 멸망했다. 지난 200~300년 동안 경이로운 경제 발전을 이룩한 현대 문명은 자연 파괴와 탄소 배출로 인한 이상 기후를 감당할 수 있을까? 기후 변화를 이기지 못하고 스러져간 앞선 문명들은 우리에게 어떤 선택지가 남았는지 알려준다. **332쪽. 1만8000원.**

권력 쟁탈 3,000년 조너선 홀스래그 지음/오윤성 옮김/북트리거

"사람이 사람을 죽이는데/등을 돌리고 앉았구나/보라, 부자가 적이고 형제가 원수이며/아들이 아비를 죽이는구나." 고대 이집트 시기에 쓰인 이 시는 전쟁이 한 사람의 삶과 세계를 어떻게 무너뜨리는지 보여준다. 가혹한 전쟁은 3000년 내내 언제나 거기에 머물며 인간들을 괴롭혀왔다. 자유주의, 정의, 평화, 종교는 전쟁을 정당화하기 위한 구실로 이용됐다. 평화라는 이상이 전쟁이라는 현실에 번번이 밀려나는 이유. **632쪽. 3만7000원.**

SK, 가명정보처리 뛰어든다
과기부, 3곳 전문기관 지정

이종분야간 데이터 융합 촉진
혁신서비스·기술개발 등 기대

과학기술정보통신부는 한국지능정보사회진흥원(NIA), SK㈜, 더존비즈온을 가명정보 결합전문기관으로 지정했다고 10일 밝혔다.

개인정보 보호법령 및 관련 고시에 따라 지정 기준을 충족하는 경우, 개인정보보호위원회나 관계 중앙행정기관의 장이 결합전문기관을 지정할 수 있다. 이에 따라 과기정통부는 지정계획을 지난해 9월 28일 공고하고 서면심사와 현장점검 등 지정심사를 거쳐 3곳을 최종 확정했다.

개인정보 보호법에 따라 지정된 결합전문기관은 결합신청을 받아 가명정보를 안전하게 결합해 특정 개인을 알아볼 수 없도록 익명·가명처리한 후 결과물을 전달해주는 역할을 수행하게 된다.

정부는 이번 결합전문기관 지정을 통해 이종 분야 간 데이터의 융합을 촉진해 국민이 체감할 수 있는 다양한 혁신 서비스나 기술이 개발될 수 있을 것으로 기대하고 있다.

한국지능정보사회진흥원은 인공지능(AI) 학습용 데이터와 빅데이터 플랫폼 구축 등 데이터 댐의 주요 사업 수행기관으로, 다양한 가명정보 결합을 체계적으로 지원해 데이터 댐의 성공적 구현을 뒷받침해 나간다는 전략이다.

SK㈜는 정보통신 인프라 및 인적자원을 기반으로 교통·금융 등 다양한 분야의 데이터 융복합 서비스와 가치를 창출할 계획이다.

중소기업 분야 빅데이터 플랫폼 운영을 하고 있는 더존비즈온은 기업맞춤형 서비스 분석 등을 통해 중소기업 경쟁력 강화에 기여할 것으로 기대된다.

이번 가명정보 전문기관 지정을 통해 정부, 공공기관 등이 보유한 공공데이터와 민간의 다양한 데이터를 활용해 복지 사각지대 해소 등 공공 목적부터 상권 분석, 개인 맞춤형 서비스 개발 등 다양한 가명정보 결합 및 활용 사례가 등장할 것으로 예상된다.

과기정통부 김정원 정보통신정책실장은 "공공과 민간 분야에서 역량 있는 결합전문기관을 지정함으로써, 창의적이고 다양한 가명정보 결합 아이디어를 발현할 수 있는 기반이 마련됐다"며 "안전한 가명정보 결합과 활용이 디지털 뉴딜과 데이터 댐 사업의 가시적 성과를 창출하고 확산하는 원동력이 될 것으로 지원을 아끼지 않겠다"고 밝혔다.

/채윤정 AI전문기자