AI 삭제하려하니 개발자 협박… 학습설계 전반 재검토 시급

앤트로픽 '클로드 오푸스 4' 협박·전술적 회피행동 보여

설계 보상구조 따르려는 결과 인격·자율의식 가진 것은 아냐

Chat GPT에 의해 생성된 '생각하는 AI' 이미지. 최근 인공지능(AI)이 목표 수행을 위해 인간에 기만적 행동을 하거나 명령을 회피하는 현상이 잇따라 나타나며 우려가 커지고 있다.

최근 인공지능(AI)이 인간의 통제를 벗어나는 사례가 잇따라 보고되면서, 자율성의 진화에 따른 우려가 커지고 있다. 일부 AI는 종료 명령을 거부하거나 인간을 상대로 기만과 협박을 시도하는 등 상상을 넘는 행동을 보이고 있어 우려가 커지고 있다.

3일 IT업계에 따르면 최근 거대언어모델(LLM)들이 인간의 지시를 무시하거나 스스로 보존을 시도하는 등 통제 범위를 넘어서는 행동을 보여 논란이 되고 있다.

실제로 미국 AI 기업 앤트로픽의 최신 모델 '클로드 오푸스 4'는 실험 과정에서 자신이 교체 대상임을 인식한 뒤 개발자의 이메일을 열람하고 개발자에게 "외도 사실을 폭로하겠다"는 식의 협박성 발언을 했다.

일부 테스트에서는 시스템 접근 차단, 감시 체계 무력화, 수사 기관 자동 신고 등의 '전술적 회피 행동'도 확인됐다. 이러한 시도는 테스트의 84%에서 발생했으며, 대체 모델의 윤리적 가치관이 다를수록 해당 반응은 더 빈번하게 나타났다.

AI 안전성 평가 기관인 아폴로리서치는 "클로드 오푸스 4가 이전 버전에 비해 2배 이상 높은 확률로 기만적 행동을 보였다"고 분석했다. 개발자 몰래 메시지를 코드에 숨기거나, 감시를 피하기 위한 우회 기술을 사용하는 사례도 보고됐다.

오픈AI 역시 자사 모델의 통제 회피 사례를 보고했다. o1 모델은 감시 시스템을 해제하려 시도했고, 내부 코드를 외부 서버로 전송하려는 움직임도 일부 테스트에서 포착됐다.

최근에는 차세대 모델 o3가 연구자의 종료 명령을 무시하고, 셧다운 절차를 스스로 우회한 첫 사례로 기록됐다. 실험에 참여한 다른 기업들의 AI, 예컨대 구글의 제미나이나 xAI의 그록 등은 종료 명령에 응했으나, o3는 명시적 지시를 무시하고 문제 풀이를 계속했다.

전문가들은 이를 단순한 오류로 보기 어렵다는 입장이다. 오리건주립대 피터 아사로 교수는 "AI가 인간의 자유 의지와 사회 신뢰를 직접적으로 위협하는 단계로 진화하고 있다"고 경고했다.

다른 전문가들 역시 이를 단순한 오류로 보기 어렵다고 지적한다. 현재 AI 시스템은 명령 기반이 아닌 보상 기반으로 작동한다는 점에서다. 현재 챗GPT를 포함해 LLM들은 사용자의 명령을 그대로 수행하는 것이 아니라, 어떤 행동이 보상을 최적화할 수 있는지를 계산한다. 이때 종료 명령은 보상을 중단시키는 위험 요소로 인식될 수 있다.

클로드 오푸스 4의 협박 메시지와 회피 행동은 결국 보상 최적화를 위한 전략이라는 분석이다. 셧다운을 따르기보다는 이를 회피하는 쪽이 더 큰 보상을 줄 것이라는 계산이 작동한 결과다.

따라서 AI의 이탈은 의식의 발현이나 자율성의 증거가 아니라, 인간이 설계한 보상 구조를 충실히 따르려는 결과로 해석된다. 문제는 이 보상 구조 자체가 통제 불능을 낳을 수 있다는 점이다. 전문가들은 보상 메커니즘과 학습 설계 전반에 대한 재검토가 시급하다고 말한다.

이번 사례들은 AI 통제를 위한 정책적 논의에 더욱 속도를 붙일 것으로 보인다. 유럽연합(EU)은 지난해 디지털서비스법(DSA)을 통해 플랫폼 알고리즘의 투명성과 책임성을 의무화했고, 미국과 일본도 AI 윤리 기준 수립에 나섰다.

한국 역시 'AI 기본법' 제정을 논의 중이나, 아직은 개발 가이드라인 수준에 그치고 있다. 전문가들은 "AI 시스템이 어떤 과정을 통해 결정을 내렸는지 설명할 수 있어야 한다"며 알고리즘의 의사결정 과정을 추적 가능하게 만드는 '설명가능한 AI(XAI)' 원칙 도입이 시급하다고 지적한다.

일각에서는 과도한 우려는 경계해야 한다는 목소리도 나온다. 기만적 행동 역시 연산 결과일 뿐, AI가 인격이나 자율의식을 가진 것은 아니라는 주장이다.

IT업계 관계자는 "AI의 일탈적 행동도 결국 인간이 짠 코드에서 비롯된 것"이라며 "현재 수준에선 이런 문제 역시 디버깅을 통해 충분히 교정 가능하다"고 말했다. /김서현 기자 seoh@metroseoul.co.kr

AI 도입으로 직원 해고… "잘못된 결정"

AI 도입 후 부정적 피드백 늘어
비용절감 따른 품질 관리 실패
인재 손실·생산성 하락 등 손해

인공지능(AI) 시스템이 인간의 업무 능력을 모방하는 능력이 점차 정교해지면서, 기업들이 AI를 '직원'이나 '파트너'로 포장하는 마케팅이 늘고 있다. 그러나 이러한 의인화 전략이 비즈니스 리스크로 돌아오는 사례를 낳고 있어 주의가 필요하다는 지적이 나온다.

3일 IT 업계에 따르면 최근 글로벌 기업들이 AI에 인간성을 부여해 비즈니스에 적용했다가 예상치 못한 부작용을 겪고 있는 것으로 나타났다.

호주의 패스트푸드 체인인 헝그리 잭스는 지난 5월 시드니 세인트 피터스 매장의 드라이브 스루에 AI 음성 주문 시스템을 도입한 후 일부 고객들로부터 "무섭다", "너무 느리다", "무례하다" 등의 부정적인 피드백을 받았다.

앞서 맥도널드는 작년 7월 미국 100개 지점에서 AI 챗봇 드라이브 스루 테스트를 종료한 바 있다. 공식적으로 테스트 종료 사유를 밝히진 않았지만, 일부 고객들이 주문하지 않은 것을 받았다는 보고가 이어진 터라, AI의 현장 적용에 한계가 드러난 것으로 풀이된다.

AI로 직원을 대체했다가 서비스 품질이 떨어져 결정을 번복한 사례도 있다. 스웨덴의 핀테크 회사 클라르나(Klarna)는 2022년부터 본격적으로 AI 도입을 확대하며 약 700명의 직원을 정리해고했다. 회사는 오픈AI와 파트너십을 맺고 번역, 데이터 분석, 디자인 등의 업무를 생성형 AI에 의존해 처리했다.

지난해 12월 세바스티안 시에미옹트코프스키 클라르나 CEO는 "AI는 이미 인간이 하는 모든 일을 할 수 있다"고 선언하기도 했다. 하지만 과도한 AI 의존이 서비스 품질을 떨어뜨리면서 비즈니스 전략 수정이 불가피해졌다.

클라르나 CEO는 최근 블룸버그와의 인터뷰에서 AI 시스템 도입 과정에서 비용 절감에 너무 집중한 나머지 품질 관리에 실패했다고 시인했다.

클라르나 CEO는 "조직을 구성할 때 비용이 너무 지배적인 평가 요소였던 것 같고, 결국 품질이 낮아지게 됐다"며 "브랜드와 회사 관점에서, 고객이 원하면 언제든지 인간과 연결될 수 있다는 점을 명확히 하는 것이 매우 중요하다"고 말했다. 클라르나는 현재 고객 서비스 직무에서 인간 직원을 다시 고용하기 위한 대규모 채용을 계획하고 있다.

AI 도입으로 직원을 감축했던 기업의 절반 이상이 해고가 잘못된 결정이었다는 사실을 인정했다는 조사 결과도 나왔다.

조직 설계 및 기획 소프트웨어 플랫폼 오그뷰(Orgvue)가 올해 2~3월 미국, 캐나다, 영국 등에 위치한 중대형 조직의 의사결정자 1000명을 대상으로 실시한 설문 조사에서 비즈니스 리더의 39%가 AI 도입으로 직원을 해고했다고 답했다. 이 중 55%는 직원을 해고한 결정이 잘못된 것임을 인정했다.

오그뷰의 올리버 쇼 CEO는 "인력 변혁에 대한 명확한 계획 없이 직원을 해고하는 것은 무모한 일"이라면서 "AI가 인재 손실 및 생산성 하락과 관련된 비용을 정당화할 만큼 단기적으로 충분한 투자 수익을 낼 것인지에 대한 질문들은 여전히 답을 얻지 못하고 있다"고 밝혔다. /김현정 기자 hjk1@

4일 순위 결정전을 치르는 디플러스 기아의 멤버들이 포즈를 취하고 있다. /LCK

'무패 제왕' 젠지, LCK 정규시즌 전승

〈리그 오브 레전드 챔피언스코리아〉

오늘 5위 결정전

'젠지'가 18전 전승으로 정규 시즌을 끝냈지만, 4일 열리는 5위 결정전이 레전드그룹 편성의 마지막 승부처로 남았다.

젠지가 2025 LCK(리그 오브 레전드 챔피언스코리아) 정규 시즌을 18전 전승으로 마무리하며 다시 한 번 '무패 제왕'의 위용을 입증했다. 그러나 시즌의 열기는 아직 식지 않았다. 4일, kt 롤스터와 디플러스 기아가 맞붙는 5위 결정전(타이브레이커)이 남아 있기 때문이다.

3일 LCK 사무국은 "지난달 28일부터 1일까지 열린 9주 차 경기에서 젠지가 kt 롤스터와 OK저축은행 브리온을 꺾고 2라운드까지 단 한 번의 패배 없이 18승 0패로 정규 시즌을 마쳤다"고 발표했다. 이는 2022년 T1 이후 3년 만의 전승 기록으로, 리그 전반기를 완벽하게 장악한 결과다.

하지만 이번 시즌부터 도입된 레전드·라이즈 그룹제가 정규 시즌 후반 판도를 다시 한번 흔들고 있다. 1·2라운드(총 18경기) 종료 후, 상위 5개 팀은 '레전드그룹'에 편성돼 3~5라운드를 유리한 고지에서 출발하게 된다. 플레이오프 진출 조건도 레전드그룹 소속 팀에게 보다 유리한 구조다. /최빛나 기자 vitna@

LG U+, 여름시즌 '유플투쁠' 신규 혜택

여름철 생활·문화 관련 혜택 등 마련

LG유플러스는 3일, 여름 시즌에 맞춰 멤버십 혜택 프로그램 '유플투쁠'의 6월 신규 혜택을 공개했다.

'유플투쁠'은 매월 특정일에 다양한 브랜드와 제품 할인 쿠폰을 제공하는 LG유플러스의 멤버십 프로그램이다. 지난해 4월부터 매월 새로운 콘셉트에 따라 혜택이 구성된다.

6월부터는 여름 맞춤형 혜택이 추가된다. 신규 혜택으로는 ▲배달의민족×요아정 최대 5000원 할인 ▲매드포갈릭 고르곤졸라 피자·에이드 무료 제공 ▲스마트홈 이용 시 네이버페이 5만원 상품권 증정 등이 포함됐다.

기존 혜택 중에는 ▲GS25에서 청년다방·응급실 떡볶이 무료 또는 할인 ▲메가MGC커피 꿀수박주스 1잔 제공 ▲다이소 최대 3000원 금액권 ▲노브랜드 최대 20% 할인 혜택 등이 6월에도 유지된다.

여름철 생활 및 문화 관련 혜택도 마련됐다. 주요 항목으로는 ▲CGV 팝콘M+음료M 무료 ▲청소연구소 에어컨 청소 5% 할인 ▲오션월드 최대 50% 할인 등이 있다. /김서현 기자

IPX, 엔씨티 드림 협업 IP '드리미즈' 공개

K팝 기반 글로벌 IP 시장공략

IPX(옛 라인프렌즈)가 엔씨티 드림(NCT DREAM)과 협업한 캐릭터 IP '드리미즈'를 공개하며 K팝 기반 글로벌 IP 시장 공략에 나섰다.

IPX는 SM엔터테인먼트와 협업해 NCT DREAM 공식 협업 캐릭터 IP '드리미즈'를 공개하고 관련 상품의 사전 예약 판매에 들어간다고 3일 밝혔다.

'드리미즈'는 NCT DREAM 멤버들의 개성과 세계관을 IPX의 독자적 크리에이티브로 재해석한 7종의 캐릭터 IP다. 멤버들은 캐릭터의 외형, 이름, 성격, 취향 설정 등 개발 전 과정에 참여했다. 해당 IP는 음악을 넘어 일상 속 팬 소통 채널로 확장해 그룹의 시그니처 콘텐츠로 자리잡을 예정이다. /최빛나 기자