129 ELIZA 효과

Eliza Effect

AI에 실제 의도·감정이 있다고 과대 해석하는 심리적 현상

- 언어적 반응만으로도 지능·감정이 있다고 오해하는 인지적 착시
- AI의 표현 방식이 인간적 의미를 불러일으켜 과신을 유발하는 효과

ELIZA 효과의 유래

출처 : Nielsen Norman Group

ELIZA 효과는 단순한 기계적 반응에도 사람처럼 생각하거나 느낀다고 착각하는 현상을 의미합니다. 1960년대 MIT의 Joseph Weizenbaum이 개발한 초기 대화 프로그램 'ELIZA'에서 비롯된 이름으로, 이 프로그램은 사용자 문장을 되풀이하거나 일부 단어를 바꿔 재구성하는 단순 규칙 기반 시스템이었습니다. 그럼에도 많은 이용자가 ELIZA가 자신을 "이해하고 공감한다"고 믿었고, 이를 계기로 인간이 언어적 표현만으로 기계에 인격과 감정을 투사하는 경향이 있다는 점이 밝혀졌습니다. 현대의 생성형 AI는 당시보다 훨씬 자연스럽게 말하고 복잡한 질문에 대응하기 때문에, ELIZA 효과는 초기보다 훨씬 강하게 나타나게 됩니다.

ELIZA 효과의 원인

ELIZA 효과는 인간의 의인화 경향과 언어에 대한 과도한 신뢰에서 주로 발생합니다. 사람은 유창한 언어 능력을 지능의 핵심으로 인식하기 때문에, 문맥에 맞는 설명이나 공감적 문구가 등장하면 그 뒤에 "의미를 이해하는 주체"가 있다고 자연스럽게 가정합니다. 생성형 AI는 감정 표현, 전문적 어조, 친근한 대화 스타일을 매우 자연스럽게 구성하기 때문에 이러한 착각은 더욱 강화됩니다. 또한 AI가 안정적 대화를 지속하면 사용자는 관계 형성이나 의도적 반응이 있다고 오해하기 쉽고, 모델의 실제 작동 방식이 통계적 패턴에 기반한 것이라는 점을 잊게 됩니다. 알고리즘의 불투명성 역시 사용자의 상상 여지를 넓혀, 실제보다 더 높은 능력과 이해력을 투사하게 만드는 요소로 작용합니다.

ELIZA 효과의 함의

ELIZA 효과는 AI와 인간의 상호작용을 이해하기 위한 핵심 개념입니다. 사람은 언어적 표현만으로도 기계에 의도·감정을 투사하는 경향이 있어, AI가 실제로 수행하는 통계적 처리와 사용자가 느끼는 지능 사이에 큰 간극이 생길 수 있습니다. 또한 ELIZA 효과는 AI 설계에서 투명성·설명 가능성·역할 표기 등이 중요한 이유를 설명하며, 의료·상담처럼 인간적 판단이 중요한 영역에서 AI 사용 기준을 마련하는 근거가 됩니다. 즉, ELIZA 효과는 AI를 인간과 동일시하지 않고 기술적 한계를 명확히 인지하기 위한 출발점으로 평가됩니다.