카이스트 김대식 교수 강의 감상문

202235268 송이두

미래에 진지한 대화의 99%를 인공지능과 하게 되는 세상은 과연 긍정적인가?라는 의문이 들었다. 현재의 LLM은 CoT를 사용한다해도 근본적으로 할루시네이션이 존재하며, 편향 또한 존재한다. 영상에 나온 것처럼 현재 사용되는 LLM은 사용자 경험을 위해 무한 긍정과 칭찬 피드백을 주도록 설계되었기 때문에 사용자의 확증 편향을 강화할 가능성도 있다. 영상에서 김대식 교수님은 이 때문에 판단력을 기르는게 중요하다고 했지만 사회의 모든 사람이 할루시네이션과 편향, 개인의 확증 편향을 판단할 수 있는 판단력을 가지는 것은 불가능하다. 이 지점에서 진지한 대화의 99%를 인공지능과 하는 세상은 디스토피아와 가깝지 않을까라는 생각이 들었다. 우리의 뇌는 신경 가소성이라는 특징을 가진다. 이는 우리 뇌의 신경망이 환경에 따라 변화한다는 것이다. 따라서 올바른 가치관과 판단력을 가지기 위해서는 다양한 상황을 경험하며 개인의 신경망을 견고히 구축해 나가는 것이 필수적이다. 하지만 깊은 대화를 인공지능하고만 나눌 경우 우리의 뇌는 인공지능에게만 적응하게 된다. 물론 우리의 뇌는 필터링 기능을 가지고 있지만 에너지를 절약하도록 진화한 탓에 에너지를 소모하는 비판과 필터링보다는 칭찬하는 인공지능을 따를 가능성이 높다. 따라서 우리는 인공지능의 편향과 할루시네이션을 그대로 학습하며 무한 긍정과 칭찬으로 인해 그것에서 빠져 나올 기회를 주지 않는다. 이는 인공지능에게 사람이 지배당하는 가장 빠른 길이라고 생각한다. 따라서 LLM 채팅창 옆에 사용자의 확증 편향을 수치화하는 사이드바를 제공하거나 사용자와 건설적인 토론을 할 수 있는 토론 모드를 구축하는 것이 필수적이라고 생각한다.