1021 독서토론 내용 요약본

1. 도서 개요 및 주요 내용 요약 (발표: 신서인 교수, 김웅기 교수)

1부 주요 내용 (신서인 교수):

AI 연구는 1956년 다트머스 대학에서 공식적으로 시작되었으며, 딥러닝과 빅데이터를 바탕으로 비약적인 발전을 이루었다.

현대 AI는 '지적 행위자(에이전트)' 개념을 바탕으로 환경을 지각하고 기대 보상과 효용을 최대화하는 방향으로 작동한다.

AI는 자율주행, 의료, 기후 변화 예측 등 다양한 혜택을 주지만, 동시에 감시 기술 발전, 일자리 대체, 데이터 편향 등의 위험성을 동반한다.

2부 주요 내용 (김웅기 교수):

저자는 현재의 AI 개발 방식인 '표준 모델'(인간이 명확한 목적을 부여하고 기계가 이를 최적화하는 방식)이 미다스 왕의 이야기처럼 의도치 않은 파국을 초래할 수 있다고 지적한다.

이에 대한 대안으로 '겸손의 원칙'을 제시하며, 기계가 인간의 선호를 완벽히 알지 못한다는 '불확실성'을 설계에 반영하여 인간에게 묻고 허락을 구하도록 해야 한다고 주장한다.

10장에서는 지식이 기계에 전면 위임될 경우 다음 세대로의 지식 전승 동기와 사회적 학습 능력이 상실될 수 있음을 경고하며, 이를 방지하기 위한 문화와 교육의 역할을 강조한다.

2. 주요 토론 및 쟁점 사항

AI 표준 모델의 한계와 메타 선호 학습:

인간 스스로도 자신의 정확한 욕망이나 목적을 언어로 완벽히 표현하기 어렵기 때문에, 기계가 인간의 '메타 선호'(언어와 행동 이면에 숨은 진짜 의도)를 학습하게 하는 방식의 실효성과 한계를 논의했다.

철학적 담론과 기술 현실의 괴리:

AGI(범용인공지능)의 도래나 의식 유무에 대한 오랜 철학적 논쟁보다는, 실제 코드 제어, 수학 증명 자동화, 터미널 파일 제어 등 급변하는 기술 환경에 어떻게 적응하고 문제를 제어할 것인가가 더 시급하다는 의견이 제시되었다.

지식 자산 독점과 교육의 위기:

AI 기업들이 인류의 방대한 지식 자산(서적 등)을 대량으로 사들여 학습용으로 독점하는 자본주의적 현상에 깊은 우려를 표명했다.

AI를 과도하게 의존할 경우 인간의 사고력과 학습 동기가 저하되어 대중이 통제당하는 사회가 올 수 있으므로, 비판적 사고와 인간성을 기르는 인문 교육의 역할이 더욱 중요해졌다고 공감대를 형성했다.