

인공지능을 어떻게 제어할 것인가?

이동재

2024년 11월 28일

Abstract

최근 인공지능의 발전으로 인해 인공지능은 메모리 접근, API 실행, 화면 조작 등 더 많은 권한을 가지게 되었다. 이러한 권한 확대는 인공지능의 행동 제어를 더욱 중요하게 만들고 있다. 본 글에서는 인공지능 제어의 필요성을 강조하고, 이를 달성하기 위한 주요 접근 방식을 논의한다.

최근 인공지능은 에이전트(Agent)라는 자율 실행 소프트웨어로 발전하고 있다. 에이전트는 환경을 탐색하며, API 호출이나 화면 조작을 통해 임무를 수행한다. 현재 이러한 에이전트는 실험 환경에서 제한적으로 실행되지만, 앞으로는 실제 환경에서 더 넓은 권한을 가지고 인간과 협력하거나 독립적으로 작동하게 될 것이다.

인공지능이 인간과 대등한 권한을 가지게 되면, 행동 제어는 필수적이다. 예를 들어, 에이전트의 실수로 중요한 데이터가 유출되거나, 잘못된 의사결정이 시스템 전체의 붕괴를 초래할 가능성이 있다. 이러한 위험은 최근 대체를 이루는 신경망 기반의 모델에서는 미리 알아내기 대단히 어렵다. 따라서 강력하고 신뢰할 수 있는 제어 메커니즘이 필요하다.

하지만, 현재 인공지능 안전성은 정렬(Alignment)이라는 방식에 의존하고 있다. 정렬은 강화 학습을 통해 인간이 원하는 규칙을 모델에 학습시키는 방식이다. 이러한 훈련 기반 방식은 항상 예측 가능한 동작을 보장하지 못한다. 따라서, 정렬만으로는 에이전트의 안전성을 확보하기 어려우며, 논리적 검증을 포함한 추가적인 접근이 필요하다.

프로그램 검증과 유사하게, 인공지능의 행동을 검증하는 방법은 정적 방식과 동적 방식으로 나눌 수 있다. 정적 방식은 인공지능 모델을 직접 분석하거나, 검증 가능한 알고리즘 형태로 변환하는 것이다. 예를 들어, 신경망 기반 언어 모델에서 학습한 규칙을 명시적이고 해석 가능한 형태로 추출하는 방법이 있다.

동적 방식은 실행 과정에서의 검증에 초점을 맞춘다. 에이전트가 생성한 결과물이 주어진 규칙을 준수하는지 실시간으로 확인하거나, 결과물의 위험성을 사전에 탐지하는 방식이 이에 해당한다. 이러한 검증 작업은 실제 환경에서 에이전트가 실행되는 동안 지속적으로 수행되어야 할 것이다.