인문대학 AI TF 독서토론
2026.10.21.(수) 14:00~15:30

스튜어트 러셀, 『어떻게 인간과 공존하는 인공지능을 만들 것인가: AI와 통제 문제』 1~5장

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control.* Viking.

발제: 신서인

1. 들어가며

이 책은 AI가 인간에게 이로운 방법으로 기능하도록 하기 위한 방법을 제안하는 것을 목적으로 한다. AI 연구는 1956년에 다트머스 대학에서 공식적으로 출범하였다. 존 매카시, 마빈 민스키, 클로드 섀넌, 너대니얼 로체스터의 연구는 인간의 학습 능력이나 지능을 기계가 모사할 수 있을 만큼 정확히 기술할 수 있다는 가정에 기반하여, 기계가 어떻게 언어를 사용하고 추상화를 이루고 개념을 형성하고 문제를 해결하며 스스로를 개선하는지 알아내려고 시도하였다. AI 연구는 초창기 기계 번역이 만족할 만한 성능을 내지 못하면서 잠시 침체기를 겪기도 하였으나 대규모 확률 추론 시스템과 심층 학습(deep learning)에 기반하여 비약적인 발전을 이루었다. 2016년 알파고와 이세돌의 대국을 계기로 인류는 AI의 위력을 확인하였으며, 그 이후 AI가 가져올 경제적·사회적 혜택에 대한 기대로 AI 연구와 산업에 엄청난 투자가 이어졌다.

AI의 급격한 발전은 기계가 인간을 정복할 것이라는 비관론으로 이어지기도 하지만, 저자는 수 년 내에는 그러한 일이 일어날 가능성이 낮다고 보고 있다. AI가 인간의 선호를 예측 가능하게 한다면 AI가 절대로 우리를 지배할 능력을 갖지 못하게 할 수 있다고 말한다. 저자는 노버트 위너의 <자동화의 몇 가지 도덕적·기술적 결과(1960)> 라는 논문에서 '우리가 기계에 부여하는 목적이 우리가 정말로 원하는 목적이 되도록 확실히 조치해야 할 것이다.'라는 문장을 인용하면서, 기계가 인간보다 더 지적인 존재가 되어 자신의 목적을 추구하도록 허용하는 대신 기계가 인간의 목적을 추구하도록 종용한다면 충분히 지적이면서도 인류에게 유익한 기계를 설계할 수 있을 것이라고 주장한다. 사실 우리도, 기계도 우리의 목적이 무엇인지 명확하게 알지 못한다. 그런데, 오히려 이 점이 해결의 실마리가 될 수 있다고 한다. 목적이 불확실하면 기계는 인간을 따를 수밖에 없다. 오히려 목적을 불확실하게 함으로써 기계가 허가를 요청하고, 수정을 받아들이고, 작동을 멈추는 일을 허용하도록 한다면 성공적인 미래를 보장할 수 있을 것이다.

2. 인간의 지능과 기계의 지능

지능은 자신이 지각한 것을 토대로 자신이 원하는 것을 성취할 수 있게 하는 것이다. 우리의 뇌는 보상 체계에 기반하여 특정한 행동을 강화한다. 대개의 생명체는 보상을 최대화하는

학습을 통해 진화해왔다. 우리는 행동하는 주체, 곧 행위자(agent)이며, 논리 추론, 의도적 계획, 지혜, 재치, 상상, 창의성 같은 지적 특징을 바탕으로 지적인 행동을 수행한다. 합리적인 행위자는 기대 효용을 최대화하는 쪽으로 행동하지만, 행위자가 다수가 되면 상황이 복잡해진다. 우리는 인류에게 혜택이 돌아오도록 상호 협력을 이루어야 할 것이다.

컴퓨터과학의 핵심 개념은 알고리듬이다. 알고리듬은 무언가를 계산하기 위해 정확하게 구체적으로 지정된 방법이다. 컴퓨터 하드웨어가 발달하면서 기억 용량이 증가하고 처리 속도가 빨라지게 되었다. 이로써 알고리듬을 더 빨리 돌리고 더 많은 정보를 다룰 수 있게 되었다. 그런데 엉성하게 설계된 알고리듬을 더 빠른 컴퓨터에 돌린다고 해서 알고리듬이 더 나아지지는 않는다. 틀린 답에 더 빨리 다다를 뿐이다. 문제는 소프트웨어이다. 그러나 우리는 기계를 진정으로 지적으로 만드는 방법을 아직 알지 못한다. 또한 현실 세계는 복잡성을 띠고 있기 때문에 기계가 지적인 존재가 되고 성능이 좋아진다고 하더라도 합리적인 결정을 내릴 수 있다는 보장이 없다.

앨런 튜링은 <계산 기계와 지능(1950)>에서 (생각이라고 묘사되어야 하겠지만[1]) 사람이 하는 생각과는 전혀 다른 무언가를 기계가 할 수도 있을 것이라고 했다. 현대 AI의 핵심 개념은 지적 행위자(agent)[2], 즉 지각하고 행동하는 무엇이다. 지적 행위자를 구축하기 위해서는 환경, 관찰과 행동, 목적이라는 세 가지 문제를 고려해야 한다. 이 중 '기계에 부여하는 목적(objectives)'은 정확하게 기술된 목표(goal)라는 형태를 취할 수 없으며 목표를 달성하는 행동 사슬 같은 것도 없다. 다만, 행동을 결정하기 위해서는 목표를 달성할 확률을 최대화하는 계획을 세워야 하고, 이때 효용 함수를 사용할 수 있다. 현대 AI는 목표와 논리 대신에 확률을 사용한다. 기계는 가능한 결과들에 각 확률을 곱하여 평균화한, 기대 효용이나 기대 보상의 총합을 최대화하는 행동을 하고자 한다.

주디어 펄은 베이즈 논리를 추론 알고리듬에 도입하였다. 베이즈 정리는 사전 확률이 어떤 증거를 관찰한 결과로써 사후 확률이 되는지를 아주 단순한 방식으로 기술하는데, 사후 확률은 새로운 사전 확률이 되므로 베이즈 갱신 과정은 무한히 반복된다. 기대 효용의 최대화라는 현대의 합리성은 베이즈 합리성이라고도 하는데, 합리적 행위자가 세계의 현재 가능한 사태와 자신의 모든 과거 경험을 토대로 미래에 관한 가설 전체의 사후 확률 분포에 접근하다고 가정하는 것이다.

AI 연구자들은 강화 학습(reinforcement learning)을 개발하였다. 강화 학습 알고리듬은 환경에서 보상 신호를 직접 경험하면서 배운다. 이를 이용하여 기계가 유능한 행동을 할 수

---

1) 기계의 '생각'을 어떻게 기술할 것인가? 클로드는 명령을 수행하는 동안 다음과 같은 메시지를 보여준다.

생각하는 중, 사색하는 중, 추적하는 중, 저울질하는 중, 곱씹는 중, 숙고 중, 사색 중, 골몰 중, 가늠 중, ...

그러나 사실 AI가 하는 일은 사실 위의 표현 중 어디에도 해당하지 않는다. AI가 작동하는 방식, 즉 지각하고 행동하는 방식을 온전히 기술하지 못하면 설명하거나 통제하는 것은 거의 불가능한 일이라고 보아야 한다.

2) 언어학에서 agent(행위주 또는 행동주)는 의도를 가지고 행위를 하는 주체를 가리키는 의미역(thematic role)이다. 최근, AI 에이전트(AI Agent)나 에이전틱 AI (Agentic AI)와 같은 표현을 사용한다. AI 에이전트는 거대언어모델(LLM)에 외부 API, 검색(RAG), 함수 호출 등의 도구를 붙여 주어진 단일 작업을 수행하는 개별 시스템을 말하고, 에이전틱 AI는 개별 에이전트를 넘어 스스로 목표를 설정하고 계획을 수립하며, 여러 에이전트가 협업해 장기적·복잡한 문제를 해결하는 아키텍처 및 방법론을 말한다. 이러한 자동화 시스템에 agent라는 용어를 사용하는 것이 온당한지 생각해 볼 필요가 있다.

있다. 한편 사례를 통해 학습하는 방법도 있다. 예를 들어 프랑스어를 영어로 번역하는 행위자를 만들기 위해서는 프랑스어에 대한 올바른 영어 쌍을 제시하면 지도 학습(supervised learning) 알고리듬이 수백 개의 층위와 수백만 개의 매개 변수를 갖춘 인공 신경망을 사용하여 규칙을 생성해낸다. AI는 기계 번역, 음성 인식, 시각 대상 인식 등의 주요한 과제를 이러한 심층 학습을 통해 수행한다.

표준 모델 알고리듬은 정해진 목적을 최적화하는 일을 한다. 즉, 예측 정확도를 최대화하고 오류를 최소화하는 것이다. 이미지에 자동으로 꼬리표를 붙이는 일은 이러한 과정에 기반하고 있지만, 학습 데이터에 편향이 있으면 구글의 사진 분류 사례에서와 같은 오류가 발생할 수 있다. 이러한 문제는 잘못 분류하였을 때의 손실을 최소화하는 방식의 접근으로는 해결이 불가능하다. 분류 오류의 비용에 불확실성이 있음을 인정하고 이에 적절히 민감하게 반응하는 학습 및 분류 알고리듬을 작성해야 할 것이다.

3. AI의 혜택

AI는 인류에게 다양한 혜택을 제공할 수 있다. 마이크, 카메라, GPS 시스템 등이 발달하면서 AI 시스템이 사람들의 일상생활에 접근할 가능성이 열리게 되었다. 자율주행차, 인공지능 개인 비서, 스마트 홈과 가정용 로봇 등이 개발되면서 일상의 편리함이 증대되었다. 또한 인공위성이 찍은 영상을 컴퓨터 시각 알고리듬이 빠른 속도로 처리하여 전세계의 데이터베이스를 생성하면서 경제 활동, 식생 변화, 동물과 인간의 이주, 기후 변화의 영향 등을 시각화하고 예측 모델을 구축할 수 있게 되었다.

그러나 인간 수준의 범용 AI를 만들기 위해서는 해결해야 할 문제들이 많다. 첫째, 언어를 이해해야 하고 언어로 표현되지 않은 상식을 알아야 한다. 둘째, 새로운 개념을 만들어낼 수 있어야 한다. 셋째, 특정한 목표를 달성할 복잡한 계획을 세울 수 있어야 한다.

이러한 문제들을 모두 돌파하고 범용 지적 시스템이 만들어진다면, 이러한 초지능 AI는 다양한 방면으로 확장이 가능하다. 지금까지 세상에 나온 모든 책 읽는 것과 같이 기계의 감각 입력을 확장할 수도 있고, 수백만 대의 로봇 제어하는 것과 같이 기계의 활동 능력의 규모를 확대할 수도 있다. 기계가 환경 재난이 일어났을 때 사상자의 수를 최소로 줄이는 행동을 수행하도록 결정을 내릴 수도 있고, 지구 온난화를 막을 정책을 제안할 수도 있다.

그러나 초지능이라고 하더라도 인간을 모델화하는 것은 쉽지 않을 것이다. 인간은 다른 사람들의 정신적, 감정적 삶을 모사할 수 있지만, 기계는 그러지 못하기 때문이다. 기계가 인간을 인간 수준으로 이해하는 일은 다른 과제들보다 훨씬 어려운 일일 것이다.

4. AI의 오용

AI의 오용에 대해서도 주의를 기울일 필요가 있다. 감시 기술이 발전하면서 우리의 모든 것이 감시의 대상이 되고 있다. 지적 정보 통합 시스템은 구매 내역, 인터넷과 SNS 이용 양상, 전자제품 사용, 통화와 문자 기록, 고용, 건강 정보 등의 데이터를 바탕으로 우리를 정확하게 파악하고 행동을 예측할 수 있다. 또한 이러한 기술은 행동을 통제하는 데 활용될 수도 있다. 가짜 정보는 정신적 안정을 위협할 것이며, 자율 무기는 개인, 지역, 국가, 세계 수준에서 인간의 안전을 위협할 것이다.

로봇은 인간의 일자리를 빼앗아 가기도 한다. 일부 국가에서는 현재 '일'이라고 부르는 대부

분을 기계가 하게 될 미래를 대비하여 전환 계획을 세우고 있는데, 보편적 기본 소득도 그 중의 하나이다. 앞으로 상품과 서비스 대부분이 기계로부터 나오게 되면 인간은 인간만이 제공할 수 있는 일이나 인간이 제공하는 것을 선호하는 서비스에 종사하게 될 것이다.

로봇의 지위를 불필요하게, 또 불합리하게 높이는 것은 인간의 지위와 존엄성을 떨어뜨릴 위험이 있다. 인간에게 영향을 미칠 결정을 내리는 데에 로봇을 활용하는 것은 상당히 위험한데 이는 알고리듬의 편향 가능성 때문이다. 오늘날 알고리듬의 편향은 데이터에서 비롯되는 경우가 많으므로 공정성을 지키기 위해서는 데이터를 보정해야 한다. 그러나 공정성을 위해 데이터를 보정하면 예측의 정확도가 떨어지는 부작용이 생기기도 하므로 둘 사이에 균형을 잡을 필요가 있다. 인간이 충분한 이해, 권한, 자율성을 간직하도록 기계를 설계한다면 그 체계의 기술 부분은 인간의 능력을 대폭 확대할 수 있을 것이다.

5. 지나치게 지적인 AI

우리는 로봇이든 외계인이든 우리보다 더 지적인 무언가를 생각할 때면 곧바로 불편함을 느낀다. 고릴라가 인간을 보고도 그렇게 느낄 것이라고 하면서 저자는 이를 '고릴라 문제'라고 부른다. 앨런 튜링은 새로운 위험이 우리를 불안하게 만들 것이라고 하면서 우리가 전략적으로 중요한 순간에 전원을 끄는 행동을 할 것이라고 하였다.

그러나 범용의 인간 수준의 AI 시스템의 개발과 보급을 금지하는 일은 쉽지 않을 것이다. AI 연구를 그만두는 것은 인류를 더 나은 문명으로 이끌 기회를 저버리는 일이기도 하기 때문이다. 또한 범용 AI 개발을 금지하더라도 도구 AI 연구가 다른 분야에 적용되어 범용 AI에 가깝게 발전하는 일도 있기 때문에 범용 AI를 금지하는 일은 현실적으로 쉽지 않다. 더 나은 AI를 만드는 것이 왜 나쁜 일이 될 수 있는지 그 이유를 이해하는 것에서 출발해야 한다.

AI 문제는 '미다스 왕의 문제'와 닮아 있다. 표준 모델은 인간이 자신의 목적을 기계에 불어넣으려는 시도라고 할 수 있는데, 인간의 진정한 목적은 올바로 완전히 정의할 수 없으므로 실패할 수밖에 없다. 미다스의 손이 모든 것을 금으로 변하게 하는 일이나 마법사 견습생의 빗자루가 물을 길어오는 일을 멈추지 않는 것을 기술 분야의 방식으로는 가치 정렬(value alignment)에 실패하는 것이라고 표현할 수 있다. 아직 파국에 이르지 않은 것은 최근까지는 지적 기계의 능력이 한정되어 있어 세상에 미치는 영향의 범위가 한정되어 있었기 때문이다.

예를 들어 기계에게 커피를 가져오라는 목적을 부여하면 기계는 갖은 수를 써서 이 임무를 수행하려고 할 것이다. 임무를 완료하기 전에 전원이 꺼지면 목적을 달성할 수 없으므로 인간이 전원을 끄는 것도 거부할 것이다. AI 시스템이 어떤 목적을 명확하게 가진다면 자신을 존속시키기 위해 선제적으로 행동할 수 있다.

기계의 지능 폭발이 일어나면 통제 문제를 해결할 시간이 없을 것이다. 가능한 시나리오는 네 가지이다. 첫째, AI 연구를 중단한다. 둘째, 고도의 AI를 개발하는 일이 위험하다는 것을 부정한다. 셋째, AI 시스템이 인간의 통제를 벗어나지 않도록 설계하여 위험을 완화한다.[3] 넷째, 포기하고 지적 기계에 미래를 맡긴다. 중단이나 포기는 이루어지기 어려운 시나리오이므로 부정과 완화에 대해 뒤에서 살펴보도록 하자.

---

3) 인간의 선호를 강화하기 위한 데이터는 어떻게 구축할 것인가? 누구의 선호를 반영할 것이며, 데이터를 어떠한 형태로 구축하고, 어떠한 방식으로 학습하게 할 것인지에 대한 구체적인 방법이 마련되어야 한다.