진화의 끝, 인공지능?

이태수*

【요약】

인공지능 기술의 개발 목표는 인간 뇌의 기능을 기계에게 성공적으로 외주하는 것이다. 이 목표를 달성하려면 뇌의 기능 중 어떤 것의 작동 기제를 알로리즘으로 포착 할 수 있는지 그 범위가 일차적으로 문제될 것이다. 특히 관심을 끄는 대목은 전통적으로 인간에게만 특유한 것으로 이해되어 왔던 핵심적인 기능까지 수행할 수 있는 소위 강한 인공지능의 출현 가능성 문제다. 또한 강한 인공지능이 제작 가능하다고 하더라도 그것의 쓸모가 무엇인지에 대한 문제도 제기될 수 있다. 그러나 그런 문제에 답을 구하는 것보다 강한 인공지능의 제작이 성공하기 이전 단계에서도 초지능이 출현할 가능성이 있다는 사실을 더 심각하게 고려해야 한다. 도덕성이 결여된 초지능의 출현은 인간에게는 큰 위험이 될 수 있기 때문이다. 이 글에서 필자는 초지능의 출현이 인류의 운명에 대해 갖게 될 심중한 함축을 조명해보면서 AI 기술 개발의 방향에 관하여 좀 더 깊이 생각해야 할 필요가 있다는 점을 강조할 것이다.

【주제어】 기술 발달, 신체 기능의 외주, 탈육화, 알고리즘, 강한 인공지능, 진화

---

* 인천대학교

** 이 논문은 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2014S1A5B8063466).

https://doi.org/[10.34162/hefins.2019..22.001]

21세기 인간 삶의 모습은 어떻게 꾸며질까? 아무나 밑도 끝도 없이 불쑥 이런 질문을 던지지는 않는다. 그것은 지난 세기 동안 급속한 과학의 발전이 인간 삶의 모습을 어지러울 정도로 변화시키는 것을 목도한 현대인이 불안감과 기대감이 뒤섞인 마음으로 과학기술이 자신들의 삶을 앞으로 또 어떻게 바꾸어 놓을지 궁금해 하면서 던지는 질문이다. 그래서 이 질문에 대한 답은 일단 과학기술의 미래로 예상할 수 있는 범위 내에서 구해야 한다. 그 중에도 인간의 삶을 가장 획기적으로 변화시킬 분야, 그래서 가장 큰 불안감과 기대감을 불러일으키는 첨단 분야에 우선적으로 주목해야 할 것은 당연하다.

첨단 분야도 여럿이지만, 우리들 대부분은 아무래도 AI 분야를 그 중 선도적인 분야로 꼽을 것이다. 비전문가들로서는 알파고나 자율자동차와 같이 특별히 인상적인 실적을 근거로 삼아 AI 분야가 첨단 중의 첨단인 기술 분야라고 생각하게 되는 것이 자연스러운 일이다. 일단 눈에 띄는 것부터 고려한 것이지만, 그렇다고 해서 그 생각이 잘못된 것은 아니다. AI 분야를 가장 선도적인 분야로 꼽는 것에는 강력한 객관적 근거가 있다. AI 및 컴퓨터 공학 기술 전부가 수학이나 논리학과 같은 형식과학에 기반을 두고 있다는 사실이 바로 그 근거다. 이 사실에는 또 따로 논의해야 할 깊은 의미가 있지만, 그것은 앞으로 이야기가 진행되면서 어느 정도 저절로 확인이 될 터이니 여기서는 간단하게 그 이유 중 현상적인 부분만을 먼저 언급해두겠다.

형식과학의 연구는 물리학, 화학, 생물학과 같은 대표적인 자연과학과 달라 단출한 장비만으로 수행될 수 있다는 장점이 있다. 이 장점은 형식과학에 기반을 둔 기술 분야에서도 그대로 유지된다. 이 장점 덕택에 Apple이나 MS의 창업자처럼 대학 중퇴자인 아웃사이더들이 자기 집 차고에 설치해 놓은 초라한 장비를 이용해 세계 시장을 석권할 수 있는 소프트웨어를 개발해 내는 신화가 가능했던 것이다. 세계 도처에서 재력, 학력의 배경 없이 내세울 것은 머리 하나뿐인 젊은 인력이 자진해서 동원되어 주는 것이 이 분야다. 그만큼 이 분야의 개발 작업은 발걸음이 빠를 수 있다. 이런 것은 다른 분야에서는 기대하기 어렵다. 다른 분야의 첨단 기술을 개발하는 과제는 대체로 수많은

박사 인력을 모아 기업과 같은 조직 속에 묶어 관리하면서 비싼 재료, 시약, 설비 등에 아낌없이 돈을 쓰는 거대과학 규모의 연구 기획에 한 부분으로 편입되어 있다. 그 때문에 과제 수행의 발걸음이 컴퓨터 분야의 개발 작업과는 비교할 수 없이 무겁다. 앞으로도 AI개발 작업이 - 작업 환경이 다르게 조성될 가능성도 없지는 않지만 - 지금까지의 소프트웨어 개발 환경의 틀 안에서 계속 수행될 수 있다면 AI는 형식과학에 기반하고 있다는 사실에서 기인하는 이 장점 덕택에 다른 분야의 선두에서 계속 달려 나갈 것이라는 예상을 할 수 있다.

형식과학에 기반하고 있다는 것은 또 다른 중요한 측면에서도 AI의 위상을 결정해준다. 형식과학은 다른 자연과학의 기초를 마련해주는 역할을 한다. 여기서 기초라는 의미는 물리학 공부에는 수학이 필수지만, 물리학 지식이 있어야 수학을 공부할 수 있는 것은 아니라는 사실을 생각해보면 얼른 이해될 것이다. 자연과학에 기반한 기술의 개발도 형식과학에 기반한 컴퓨터 분야의 도움을 받지 않으면 제대로 수행될 수 없다. 물론 양자 컴퓨팅이나 신경망 회로의 착상처럼 물리학이나 생물학 쪽에서 도움이 올 수도 있으나, 그 때문에 전체적인 의존관계가 뒤바뀌는 것은 아니다. AI가 모든 과학 기술 분야를 선도하는 역할을 해왔고 앞으로도 그러리라는 것은 의심할 여지가 없다.

AI의 역할이 정말 그러하다면 인류의 미래는 사실상 AI 기술 개발이 어떤 성과를 낼지에 달려있다고 보아야 한다. 그런데 현재 여러 곳에서 AI 간판을 걸고 추진하고 있는 개발 기획만 시야에 담고서 AI가 꾸며 줄 미래의 그림을 그리기는 어렵다. 착상이든 공상이든 아직까지는 머릿속에만 들어 있는 것들 그리고 아예 머리에도 떠오르지 않은 훨씬 더 많은 것들이 AI가 해 낼 수 있는 일의 범위에 들어 갈 것이기 때문이다. 그러나 그런 일까지 망라하여 지능이 할 수 있는 일 전체의 일람표를 만들지는 못한다. AI 기술이 모사하려는 원본인 인간 지능이 계속 진화를 하는 한은 그것이 할 수 있는 일의 범위가 계속 커질 것이기 때문이다. 지능의 전모를 확인하는 것은 진화의 종점에서나 가능한 일일 터이니 진화의 여정 중에서 미리 그것을 다 알아낼

수는 없다. 다시 말해 진화의 여정 중에 있는 우리는 지능이 할 수 있는 일의 내용을 모두 일람표에 담아 명확한 표적처럼 세워놓고 그것을 겨냥하듯이 AI개발기획을 추진하지는 못한다는 것이다.

진화의 종점에 도달한 지능을 모사한다는 것은 원칙적으로 불가능한 기획이니까 그런 욕심을 버리는 것은 당연하겠다. 하지만 그 대신 진화 여정 중에 있는 인간의 지능을 표적으로 삼는 것까지 불가능한 것은 아니지 않겠는가 하는 생각을 할 수 있다. 실제로 현대적인 범용 컴퓨터의 기본 구조를 개념화한 튜링 기계를 구상한 튜링이 제안했던 튜링 테스트의 통과 기준이 바로 그런 생각을 담고 있는 것으로 볼 수 있다. 튜링 테스트는 튜링이 구상한 구조의 기계가 얼마나 인간의 지능을 잘 모사할 수 있는지를 평가할 목적으로 설계된 것이다. 그런데 튜링은 테스트를 받는 기계가 모사할 모본인 인간의 지능이 무엇인지 명확하게 정의하지 않았다. 통과 기준으로 규정된 것의 내용은 그냥 판정관 역할을 맡은 사람들이 기계와 대화를 나눈 뒤 대화 상대방이 기계인지 사람인지 구별하지 못하면 테스트를 통과한 것으로 간주한다 즉 지능적인 사고를 하는 기계로 인정한다는 것뿐이다. 그러니까 이 통과 기준은 AI 개발 기획이 조준해야 표적을 명확하게 설정했다기보다 개발 노력이 지향하는 바가 어떤 것인지만을 시사하고 있는 것이다.

AI 전문가들 중 일부는 아예 튜링 테스트 통과를 일차적인 개발 목표로 설정하고 연구를 진행하기도 한다. 얼마 전에는 구스트만Goostman이라 불리는 프로그램이 드디어 튜링 테스트를 통과했다고 크게 선전되기도 했다. 그러나 그 선전에 대한 사람들의 반응은 시큰둥했다. 구스트만이 통과했다는 테스트 자체가 튜링 테스트의 본래 취지에 충실하게 맞추어 설계된 것이라는 믿음을 주지 못했기 때문이다. 튜링 테스트를 실제로 시행하기 위해 챗봇 chatbot 경연대회 방식으로 게임 규칙을 구체화한 것이 설계의 핵심인데, 바로 그 대목에서 미심쩍은 생각이 들 수밖에 없었던 것이다. 우선 챗봇 경연대회에서는 시험관의 까다로운 질문을 어물쩍 넘길 수 있는 상투적인 대화 요령의 모듈을 장착한 프로그램이 유리한 점수를 얻을 수 있는데, 그런

모듈의 사용이 곧 지능적 사고의 증거라고 보기는 어렵다. 그리고 무엇보다도 대화를 실제로 진행하기 위해서는 불가피하게 시험관 수나 대화시간을 구체적으로 그러나 임의적으로 정할 수밖에 없는 것이 문제다. 시험관 수를 3명 또는 30명으로 정하든, 대화시간을 2분 또는 20분으로 정하든 그렇게 해서 얻은 결과를 근거로 사람과 같은 사고를 하는 기계가 나타났다는 주장을 할 수도 없다. 사실은 챗봇 경연대회에서는 얼마나 훌륭한 챗봇이 등장했는지 그 이상은 말할 수 없을 것이다.[1] 챗봇을 제작하는 일은 AI분야에 속하는 한 가지 과제일 뿐이지 전체를 대표하는 비중을 가진 것이 아니다. 애당초 챗봇 경연대회 용으로 설계된 테스트가 AI개발이 지향하는 바를 구현한 최종적 결과를 가려낼 수 있으리라고 기대한 AI 전문가들은 별로 많지 않았을 것이다.

사실은 튜링 테스트를 어떤 방식으로든 시행하여 일거에 AI개발의 최종적 목표를 달성했음을 확인하겠다는 것이 지니치게 성급한 욕심이다. 그것은

---

1) 테스트 방식을 구체화해서 실제로 시행되는 테스트의 의미를 전적으로 부정할 수는 없다. 그런 테스트에 대한 평가는 본래 튜링 테스트의 취지를 어떻게 이해하는가에 따라 달라질 수 있다. 튜링 테스트의 취지는 원칙적으로 두 가지 다른 관점에서 해석이 될 수 있다(I. Berkeley & C. Rice. 2018). 그 하나는 존재론적ontological 관점이라 할 수 있는 것으로 튜링 테스트를 통과한 AI가 정말 사고를 하는 존재인가라는 문제를 중심으로 테스트를 평가하는 것이다. 이 평가를 두고 벌어진 논란은 J. Searle (1981)의 유명한 '중국인 방' 논변이 제기된 이래 지금까지 계속되고 있다. 다른 하나로는 '사고한다think'는 말의 의미와 관련된 문제를 중심으로 테스트의 취지를 해석하는 의미론적semantic 관점이 있을 수 있다. 실제 시행되게끔 설계된 테스트는 행태론behaviorist적 입장을 전제하고 있기 때문에 일단 존재론적 관점에서 제기될 수 있는 문제와는 별 상관이 없고 오직 의미론적 관점에서만 평가 대상이 될 수 있다. 행태론적 입장을 취하면 사람들이 사용하는 '사고한다'는 말의 의미가 변화할 수 있기 때문에 튜링 테스트의 통과 기준도 같이 변화할 것이라는 점을 고려해야 한다. 튜링 테스트를 구상했을 때 사람들이 '사고한다'라는 말의 의미를 이해한 것보다 같은 말에 대한 후대 사람들의 의미 이해는 더 넓어질 수 있다. 아마 이 점을 염두에 두고 튜링 자신도 21세기가 되면 기계지능이 사고한다는 것을 인정하는 사람들의 수가 늘어나리라는 요지의 예언(Turing, 1950)을 했을 것이다. 구스트만이 튜링 테스트를 통과했다는 주장을 한 경연 대회 관련자도 튜링의 그 예언을 자신의 주장을 지지해주는 근거로 이해했을 가능성이 있다.

자신이 현재 가지고 있는 지식을 다 조합해서 단번에 인간을 탄생시키겠다는 프랑켄슈타인 박사의 욕심을 연상시킨다. 현재 AI 개발자들은 인간 뇌의 기능 중 일부를 모사하기 위해 그 기능을 수행할 수 있는 알고리즘을 찾는 일에 몰두하고 있다. 그렇게 지능의 일부를 계속 모사하다 보면 언젠가 인간과 거의 구별 안 되는 수준까지 모사할 수 있으리라는 기대가 AI 연구와 개발을 추동해주고 있는 것이다. 그것이 어떤 길을 거쳐 어디에 이를지 현재는 아주 희미하게만 시야에 잡힐 뿐이다. 아직 지능의 부분을 넘어선 전체의 모습은 선명하게 보이지는 않는다. 조준해볼 만큼 선명한 과녁이 보이지 않으니 우리가 할 수 있는 것은 그것이 위치한 곳을 가리키는 방향에 관해 이야기하는 것뿐이다.

그 이야기를 하기 위해 이제 나는 기술 발달 역사의 출발점까지 멀리 거슬러 올라가고자 한다. 처칠은 멀리 뒤돌아보면 멀리 앞을 내다볼 수 있다는 말을 했다고 하는데, 거기에 앞을 '더 밝게' 내다볼 것이라는 말도 덧붙일 수 있을 것 같다. 어쨌든 나는 약 200만년을 거슬러 올라가면 기술 발달 역사의 시작단계에서 AI의 앞길을 포함한 미래의 역사 전체를 좀 더 밝게 조명해줄 중요한 단서가 발견된다고 생각한다.

[그림] Oldowan-tradition stone chopper
출처: "Oldowan-tradition stone chopper", https://en.wikipedia.org/commons/oldowan.

옆의 그림이 보여주는 것이 바로 내가 생각하는 그 단서다. 이것은 올도완Oldowan도구라는 이름으로 알려진 인류가 사용한 가장 원시적인 수준의 타제석기인 찍개chopper다. 찍개는 호모 속屬의 첫 주자인 인간 종種 호모 하빌리스homo habilis가 사용한 것으로 여러 다른 모양의 샘플이 있지만, 이 그림이 보여주는 샘플은 주로 먹이를 잘게 찢는 데에 쓰였을 것으로 추측되는 모양을 하고 있다. 송곳니를 연상시키는 이 샘플의 모양은 내게 얼른 그런 추측을 하게 했다. 나는 이 찍개를 손에 쥐고 휘두르는 호모 하빌리스의 모습을 그려보면서

인류의 기술이 지닌 의미를 새삼 반추해보게 되었다.

그리고 나는 신체 기능의 외주外注가 그 핵심적인 의미라는 생각을 했다. '외주'는 이 경우 물론 비유적 표현이다. 그렇지만 신체의 일부인 송곳니가 하는 기능을 찍개에게 맡겨 대신 수행하게 했다는 대목을 지시하기 위해서는 '외주'라는 표현만큼 적합한 것이 없다. 찍개의 사용에서 발생하는 효과도 외주와 본질적으로 같다. 송곳니가 잇몸에 박혀 신체의 일부를 이루고 있을 때 그것의 공간적 사용 반경은 일단 입안으로 한정된다. 그러나 호모 하빌리스가 손에 쥔 찍개의 사용 반경은 팔을 휘두르는 만큼 커진다. 그리고 송곳니가 지닌 크기의 한계도 넘어선다. 길이 8cm정도의 저런 물건은 신장 120~30cm의 호모하빌리스의 입안에 맞는 크기가 아니다. 게다가 돌의 강도를 지녔기 때문에 단단한 것을 잘못 씹다가 훼손될 수 있는 송곳니와 달리 마음 놓고 쓸 수 있다. 쓰다가 파손이 되면 곧 다른 것으로 대체할 수 있다. 이 또한 내 몸의 일부인 송곳니에게서는 전혀 기대할 수 없는 것이다. 한번 자연이 준 신체의 주요 부분은 이미 고등 동물이 된 호모 하빌리스에게는 자연 재생이 불가능한 일회용이 되어 있었다. 한번 훼손되면 그 부분에 관한 한 그냥 불구의 상태로 지내야 한다. 외주의 효과를 이제 더 길게 말할 필요는 없겠다. 외주를 통해 호모 하빌리스는 자기 몸의 일부가 맡은 기능을 확대해서 수행하고 그 결실을 누릴 수 있게끔 되었고 수행 과정에서 겪을 수 있는 위험도 최소화할 수 있었던 것이다. 자신이 해야 할 일을 최소한의 대가만 지불하고 그 일을 더 잘 할 수 있는 남에게 맡기고 그 결실은 독식하면서도 수행과정에 발생한 피해는 수행 당사자만 겪게 하는 것, 그런 최선의 외주를 아무 도덕적 부담 없이 하게끔 해주는 것이 곧 기술이다. 우리가 지금 보고 있는 200만년이나 된 선사의 유물은 그런 외주의 형상물이다.

신체 기능을 외주하는 기술 발달의 역사는 진화의 과정에서 호모 하빌리스의 뒤를 이어 출현한 다른 인간 종에게도 계속 이어졌다. 그것은 대충 뇌가 커지는 것에 비례하는 역사로도 볼 수 있다. 특히 호모 하빌리스의 두 배가 넘는 크기의 뇌를 지닌 호모 사피엔스가 출현한 이후 기술 발달은 가속도가

붙으면서 중요한 신체 기능의 거의 전부가 외주되기에 이르렀다. 그러나 그 외주의 양상은 더 이상 찍개의 경우처럼 문외한도 직관적으로 파악할 수 있을 만큼 투명하고 단순한 것에 머물러 있지 않았다. 호모 사피엔스는 먹거리를 직접 처리할 송곳니의 역할이 절실한 야생적 현장성이 느껴지지 않는 대뇌 의존의 생존전략을 채택했다. 그 때문에 한 때는 거의 멸종 직전의 위기까지 맞았지만, 어쨌든 그 위기를 넘기고 나서 외주는 계속 복잡한 성격을 띠게 되었다.

가령 기술사의 획기적인 발명품인 바퀴의 경우를 생각해보자. 바퀴는 틀림없이 신체의 이동을 담당하는 다리의 기능을 외주 받아 제작된 것이지만 그 기하학적인 형태는 다리와 조금도 닮지 않았다. 그런 형태는 신체적인 이동 기능을 신체로부터 확실하게 추상해냈기 때문에 고안 가능했던 것이다. 이렇게 인간의 몸을 떠난 기능이 형태에 있어서도 생물학적 바탕의 아무런 흔적을 지니지 않을 정도가 되면, 인간 몸의 본래적인 환경이 아닌 하늘이나 바다에서도 그 기능이 수행되는 데에 원칙적인 지장이 없다. 인간을 태우고 하늘을 나는 비행기의 형태는 날개 부분이 새의 날개를 그저 멀리 닮았을 뿐, 인간 몸의 어떤 부분도 닮지 않았다. 물 위를 달리는 배의 돛은 인간 몸은 물론 그 어떤 생물체의 몸 형태도 연상시키지 않는다. 공간이동을 담당한 다리의 기능을 확대하는 방향으로 발전한 기술은 자연이 인간에게 준 몸의 바탕을 떠나면서 그야말로 생물 세계의 한계를 벗어나는 운송의 가능성을 실현해 보인 것이다.

이 예에서 볼 수 있는 패턴은 거의 모든 분야의 기술에 공통된 것이다. 이 패턴에서 가장 주목해야 될 대목은 탈육화脫肉化의 과정이다. 찍개와 같은 도구에 신체 기능을 외주하기 시작한 단계에서부터 기술의 개발이 지향하는 바가 탈육화라는 것은 명백했다. 그 뒤를 이은 다양한 기술의 개발은 인간의 몸 밖으로 수고할 일을 맡긴다는 외주의 본 뜻에 더 확실하게 맞추어 육신적인 특징의 한계까지 벗어나는 방향으로 철저하게 진행되었다. 그 때문에 외주의 양상이 복잡해지기는 했으나 탈육화의 지향성은 더욱 뚜렷하게 확인된다. AI의

개발도 그 방향으로 나가고 있는 것은 의심의 여지가 없다.

이제 AI 기술 개발의 방향에 관한 이야기로 넘어가기 전에 동력 생산 기술의 예를 하나 더 언급하겠다. 기술 발달의 역사에서 이루어진 발명의 성과는 어느 것 하나 뺄 것 없이 거의 모두가 다 나름 획기적인 것이었다고 할 수 있다. 발명이 이루어지고 난 뒤에나 그런 것도 가능했구나 하면서 인간의 지혜에 새삼 경탄하게 되는 그런 성과들이다. 그런 것들 중에도 동력 생산 기술은 특별히 경탄을 자아내게 하는 것이다. 그 기술을 통해 인간이 외주한 것은 인간의 신체기능 중 가장 기본적인 것이라 할 수 있는 동력 생산체제다. 그 일은 어떤 신체 부위 하나가 떠맡은 것이 아니다. 음식물 섭취에서 시작해 소화, 분해하고 운동 에너지를 공급할 수 있는 소재로 변형시켜 필요시에 연소하기까지 일련의 신체적 장치가 복합 체제를 구성해서 처리하는 것이 그 일이다. 인공적으로 동력을 생산해내는 기술은 그처럼 관찰하기도 쉽지 않은 생물학적 복합체제의 기능을 모사하는 방식으로 개발되지는 않았다. 산업혁명을 이끈 증기기관이나 발전기는 신체가 아닌 물리적 자연 현상의 관찰을 통해 얻은 착상이 개발의 시발점이 되었다. 기술이 세련되어 가는 과정에서도 신체의 작동 방식에 대한 직관적인 참조가 아닌 과학 이론의 응용이 결정적인 기여를 했다. 그와 같은 개발의 이력에 어울리게 현재 동력 생산 기술은 인간 신체의 힘이 지닌 한계를 초월하는 엄청난 자연의 위력을 관리하고 사용하는 단계에 도달해 있다. 그런 점 때문에 동력 생산 기술을 인간 신체 기능의 연장선상에 위치시키기는 어렵다. 그렇다고 해서 그 기술 역시 인간 신체 기능의 외주라는 사실이 부인되는 것은 아니다.

기술의 발달은 육체 노동의 효율을 계속 높여주면서 노동의 수고도 덜어주었지만, 인간은 힘든 육체 노동의 수고에서 완전히 벗어날 수 있다는 생각은 하지 못했다. 예컨대 농기구를 사용하면 농사일이 수월해지지만 그래도 어쨌든 농기구를 사용하기 위해서는 근육의 힘을 써야 했다. 바로 그 육체의 힘을 생산해내는 일은 산업혁명 이전까지는 외주하지 못했던 것이다. 기중기나 지렛대의 예에서 확인할 수 있듯이 그 힘을 확대할 수는 있었다. 그런

기구들은 시력을 높이기 위해 쓰는 안경과 같은 성격을 지닌 것들이다. 안경이 시력을 만들어 내는 것이 아니듯 그것들도 근육의 힘 자체를 만들어 내는 것은 아니다. 그와는 달리 증기기관이나 발전기는 근육의 힘을 확대하는 것이 아니라 아예 몸 밖에서 따로 생산해낸다. 생물체에서 그 힘은 신체의 여러 기능이 분화되기 이전 단계에서부터 갖춰져 있는 원초적인 것이고 또 여타의 특정한 기능이 수행될 때 필수적으로 쓰이는 기본적인 것이기도 하다. 그런 힘을 인공적으로 만들어낸다는 점에서 동력 생산 기술은 그때까지 꼭 필요했던 육체노동의 수고로부터도 인간을 벗어나게 해 줄 가능성을 열어놓았다고 할 수 있다.

보통 우리들은 인간이 하는 일을 정신적 활동과 육체적 노동으로 구분하기도 하는데, 그 중 육체적 노동에 관한 한 동력 생산 기술의 출현이 탈육화의 대단원에 해당하는 것일 수 있다. 대단원에 도달하면 신체 기능 중 외주할 것이 더 이상 없을 것이다. 그러나 우리에게 익숙한 정신적 활동과 육체적 노동의 구분은 아주 정확한 것은 아니다. 우리가 명확히 정의하지 않고 대충 정신적 활동이라고 부르는 것도 사실은 신체의 한 부분인 뇌가 하는 일이다. 뇌도 작동하는 데에는 에너지가 필요하다. 그것이 작동하는 것이 근육의 힘을 쓰는 것과는 다르지만 수고스럽기는 마찬가지다. 따라서 뇌가 하는 일도 외주할 생각을 할 수 있다. AI는 인간 뇌의 기능을 외주 받아 수행하는 기술의 작품이다. 그리고 탈육화는 AI기술을 통해 한 걸음 더 진전할 수 있다.[2] 뇌는 인간이 하는 일을 모두 총괄하여 수행하는 것이기 때문에 그

2) 공학 기술 발달이 탈육화 현상과 짝을 하는 까닭은 인간 뇌의 기능이 인간의 신체를 떠나서도 수행될 수 있다는 데에 있다. 따라서 H. Dreyfus (1992) 처럼 지능의 기능이 완전히 인간 신체와 떨어질 수 없다는 입장을 취하게 되면, 탈육화는 일정한 한계 내에서나 가능한 일이거나 아니면 아예 미망에 불과한 것이라고 해야 한다. 그러나 대세는 탈육화의 현상을 AI를 포함해 컴퓨터 공학 기술 발달과 짝을 지어 이야기하는 쪽으로 기울어 있다. 단, 여기서 말하는 탈육화는 한 개체로서 인간 각자의 신체와 떨어질 수 있다는 것을 의미하는 것이지 물질적 기반 자체를 떠난다는 것, 즉 비물질화dematerialization까지 의미하는 것은 아니라는 점은 잊지 말아야 한다. 컴퓨터 공학 기술의 발달과 같이 진행되는 탈육화 현상

기능을 외주하는 것이 기술 발달 역사의 진정한 대단원에 해당한다. 그러나 역사적으로 뇌 기능의 외주는 일단 인간의 몸 전체가 하는 일의 관리 보다는 뇌에만 특유한 기능이라고 생각되는 것을 맡기면서 시작되었다. 사람들이 특별히 뇌가 하는 일이라고 생각한 것은 손발이나 오관과 같은 특정한 신체 부위를 동원하지 않고, 흔히 하는 말로 머리만 써서 하는 일들이다. 가령 암산 같은 것이 한 대표적인 예다. 그런 예를 일일이 다 열거할 필요는 없겠고 여기서는 그냥 막연하게 사고think라고 불리는 것들이 그것이라고 이해하면 되겠다.

서양에서는 고대 그리스인들이 사고가 무엇인지 그 정체를 밝히려는 노력을 시작했다. 그들은 사고의 내용이든 사고의 작동 방식이든 명확히 구별하지 않고 그냥 묶어서 로고스라는 말로 그 정체를 지칭했다. 그들의 로고스에 대한 탐구는 수학과 논리학의 체계를 구상하는 결실을 맺었고 그 이래 수학과 논리학은 로고스에 대한 앎을 대표하는 학문으로 간주되었다. 컴퓨터 기술도 바로 이 두 학문을 기반으로 하는 로고스의 기술이라 할 수 있다. 컴퓨터 기술의 출발점은 일단 수치 계산의 알고리즘을 기계가 구현하게끔 하려는 시도였었는데, 단순한 수치 계산이라도 어쨌든 머리로만 하는 로고스의 일인 것이다. 그 뒤 계산기 수준을 넘어서 현대적인 범용컴퓨터로 발전하는 단계에서는 논리학이 밝혀 준 알고리즘이 컴퓨터의 핵심부에 해당하는 위상을 갖게 되었다. 한때는 논리학이 수학적 사고를 포함해 사고 일반의 기본 법칙을 밝혀 주는 학문으로 인식되기도 했다. 다시 말해 논리학이 인간지능의 정체를 밝혀주는 학문으로 인정된 것이다. 논리학의 위상을 그렇게 평가하는 사람들에게는 논리학의 형상물과도 같은 논리 회로를 중앙처리장치로 장착한 컴퓨터가 인간 뇌의 가장 중요한 기능, 즉 인간을 이성적 동물이라고 부를 수 있는 근거가 되는 기능을 수행하는 궁극의 발명품으로 보였을 것이다.

을 심층 분석한 대표적인 학자로는 누구보다도 K. Hayles (1999)을 꼽을 수 있는데, 그녀는 탈육화를 가끔 비물질화와 같은 것으로 취급하는 듯한 언명을 한다. 하지만 그녀는 컴퓨터 기술을 접하는 우리의 현상적 체험을 기술하는 맥락에서 그런 언명을 한 것이지 소프트웨어인 AI가 하드웨어없이 기능할 수 있다는 주장을 편 것은 아니다.

지금도 연역적 체계성을 핵심으로 하는 고전적인 형식 논리학의 중요성은 여전히 높이 평가되고 있다. 그러나 그런 논리학이 다루는 법칙성만이 곧 인간 지능 또는 지성의 전부라고 믿는 사람들은 많지 않다. 고전적 논리학의 법칙만으로는 인간의 의식 표층 아래에서 수행되는 사고는 물론이거니와 표층부에서 수행되는 사고의 전모 조차도 밝히지 못한다고 생각하는 사람들이 압도적으로 많아졌다. 지능 또는 지성의 경계선이 단번에 알아볼 수 있게끔 고정되어 있지 않으며 사고의 범위는 느슨할 정도로 아주 넓게 잡는 것이 합당하다는 생각이 대세가 된 것이다. 컴퓨터 분야의 연구도 대세에 맞추어 변화하였다. 지난 세기 후반부에는 컴퓨터 분야의 연구가들도 고등 수학의 문제를 푸는 것처럼 명확하게 규정된 과제를 처리하는 것만이 자신들에게 주어진 과제의 전부는 아니라고 생각하기 시작했다. 그런 과제 못지않게 풀어야 할 구체적인 과제가 무엇인지 제대로 파악하기도 어려운 불확실한 상황에서 취해야 할 행동 패턴도 지능적 사고가 마련해주는 것이라고 생각해서 적극적으로 그 방면의 연구에 나선 것이다. 그리고 나름 주목할 성과를 내기도 했다. 전통적으로 이성과 상관이 없는 것 또는 비이성적인 것으로 여겼던 어스름한 곳을 조명하고 거기서 작동하는 사고 기제의 알고리즘을 찾아내는 일이 컴퓨터 분야의 본격적인 연구과제가 되었다. 요즈음 컴퓨터를 공부하는 사람들 대부분은 고전 논리학의 추론 규칙 못지않게 베이지안 Bayesian의 추론 방식을 익혀야 하고 학습 알고리즘에 관한 수업쯤은 들어야 시대에 뒤지지 않는다는 생각을 하고 있을 것이다.

컴퓨터 분야 연구의 그와 같은 방향 전환은 곧 AI를 향한 방향 전환이기도 하다. AI 기술이 그 이전 단계의 컴퓨터 기술과 구별되는 특징은 다른 무엇보다도 지능의 개념을 아주 넓게 이해하고 있다는 데에 있다. 그 이전 단계에서는 뇌의 기능 중 일부에만 시선을 집중했다면, AI의 단계에서는 일단 뇌의 모든 기능을 시야에 담으려 드는 것이다. 이와 같은 변화는 컴퓨터 연구가들 사이에서 계산주의보다는 연결주의의 입장을 취하는 쪽의 목소리가 더 커지고 신경망 회로의 효용에 더 큰 기대를 거는 분위기에서 여실히 반영되고

있다. 동시에 AI 기술의 개발이 일층 더 높은 수준의 탈육화를 지향하다는 것도 확인된다. 로고스라고 불리는 부분에 해당하는 기능을 맡는 뇌 역시 신체의 한 부분임에도 전통적으로 그런 사실은 주목을 받지 못했다. 로고스라면 전형적인 정신적 활동으로, 즉 육체적인 것의 대립항으로 의식되었다. 따라서 로고스의 작동 기제를 담은 알고리즘을 찾아내 기계에서 구현하도록 노력하는 작업에서는 특별히 탈육화의 방향성이 두드러져 보이지 않는다. 애당초 몸의 냄새가 나지 않던 부분을 다루었던 것이다. 그러나 뇌가 처리하더라도 생명체의 몸에 직접 관련된 부분들의 경우에는 문제가 다르다. 애당초 몸 냄새가 진하게 나는 부분에서 작동 기제의 알고리즘만을 따로 떼어내 - 즉 추상해내 - 단백질 기반의 몸이 아닌 실리콘 기반의 물체에 그것을 그대로 옮기는 일은 탈육화의 전형을 확실하게 시연해 보이는 것이다.

이 대목에서 우리는 인류의 미래에 관련된 결정적인 질문에 부딪치게 된다. 뇌의 기능까지 외주를 하고 나면 더 이상 인간의 몸에는 외주할 것이 남아 있지 않게 된다. 뇌의 기능을 외주하는 것은 마지막 외주다. 하지만 뇌의 기능을 어디까지 외주할 것인가의 문제가 남아 있다. 뇌의 기능 모두를 시야에 담더라도 문자 그대로 그 전부를 외주한다는 것은 불가능한 일이다. 수십억 인간 개체의 뇌, 갓난아기들의 뇌, 뇌종양이나 알츠하이머 환자들의 뇌, 그리고 앞으로 태어날 인간 개체의 뇌가 하는 일을 다 탐사해서 뇌가 하는 일을 AI에게 외주할 수는 없다. 어차피 정상적인 뇌라는 표본을 놓고 그 기능만을 고려 범위에 넣어야 한다는 것은 두말할 필요가 없다. 표본을 설정하면서 뇌가 하는 일 중 그렇게 일부만을 선별하는 것은 당연하다. 그런 의미에서 그 표본은 일종의 추상체라고 해야 한다. 사실 과거의 기술 개발 역사에서도 그와 같은 추상화는 필수적이었다. 바퀴는 다리가 하는 일 모두를 떠맡지 않았다. 가령 태권도의 발 공격이나 공을 걷어차는 기능은 그냥 몸에 남겨두고 이동 기능만을 추상해서 구현했다. 어쨌든 뇌의 경우 되도록 많은 기능을 선별해내 표본을 설정하겠지만, 일단 소위 정상이라는 기준은 충족시켜야 선별될 것임은 당연하다. AI 연구는 또 그 표본 내에서 개발 과제를

선별해야 한다. 선별의 기준은 일단 문제의 기능을 알고리즘에 담을 수 있는가 여부일 것이다. 원칙적으로 알고리즘으로 코드화하는 것이 불가능한 기능은 AI에게 외주하지 못한다. 그 다음 중요한 선별 기준은 기술을 사용하는 주체인 인간에게 도움을 주는가 여부이겠다. 도움이 되지 않는 기능의 경우는 알고리즘으로 코드화할 가능성이 있다고 해도 AI에게 외주할 이유가 없는 것이다.

AI의 미래를 궁금해 하는 사람들은 AI가 앞으로 뇌의 기능을 모두 다 맡아서 수행하리라는 것을 당연한 듯이 전제하고 풀어내는 이야기에 특별히 흥미를 느끼는 경향이 있다. 소위 특이점singularity 그리고 강한 인공지능Strong AI 또는 범용 인공지능AGI의 출현을 주제로 한 이야기가 그런 것이다. 미래 담론으로서는 그 중 공상과학소설에 가까운 이야기들이다. 흥미로울 것은 당연하나 아직까지 AI 기술 개발의 실제는 그와는 먼 거리에 있다.[3] 정말 인간처럼 감각 또는 느낌도 가질 수 있는 존재sentient being로서 의식이 있는conscious 주체 노릇을 할 AI의 제작에 착수하여 성과를 내고 있는 기업이나 연구소가 있다는 이야기는 아직 듣지 못했다. 의식 consciousness, 일인칭적 감각qualia, 감정emotion 등 인간을 인간이게끔 해주는 특징적인 것들은 그 자체가 아닌 그 작동 구조만을 규명해내 알고리즘으로 포착하는 연구가 일각에서 진행되고 있을 뿐이다.[4] 그 연구의 가시적 성과로는 감정로봇이라 불리는 제품이 있기는 하나, 그것은 감정을 느끼는 로봇이 아니라 감정을 느끼는 인간의 행태를 어설프게 모방하거나 인간의 감정을 대충 읽을 수(?) 있는 아직은 어설픈 기계일 뿐이다. 어쨌든 현재는 의식이나 감각, 감정 등은 창발적創發的emergent 성격을 지닌 것이어서 그것 자체를 생성하거나

3) AI의 미래에 대한 낙관론자들은 AGI의 출현 가능성을 당연한 것으로 생각하고 있으며 출현 시기도 매우 앞당겨서 금세기 전반부(R. Kurzweil, 2005) 또는 아무리 늦어도 금세기가 지나기 전일 것이라고 예상한다.

4) AI 연구자들의 감정이나 의식의 문제에 관한 접근 방식과 연구 현황에 관해서는 M. Scheutz (2014)를 참조할 것. 이 요약 보고가 나온 이래 지금까지 해당 분야에서 아주 큰 변화가 있었다는 소식은 아직 없다.

학습하는 프로그램을 직접 개발하려 들지 않아도 어느 단계에 이르면 AI가 저절로 그런 기능을 수행하는 능력을 획득할 가능성이 있다는 정도의 이야기가 나오고 있는 것뿐이다.

그렇지만 인간의 상상력은 인간 뇌가 하는 일의 알고리즘이 완전히 포착된 미래로 그냥 달려 나간다. 그때가 되면 그 알고리즘을 계속 업그레이드해서 그 결과를 클라우드와 같은 저장소에 업로드하고 거기서 다른 기계에게 다운로드하는 방식으로 인간의 지능을 능가하는 지능을 지닌 AI들을 제작해 낼 수 있으리라는 상상도 하게 된다. 그 상상 속의 행복한 결말은 인간이 하는 일을 인간보다 더 유능한 AI들이 모두 맡아서 처리하고 인간은 더 이상 세상 사는 수고를 하지 않아도 되는 천국의 실현이다.

이야기가 그쯤에 이르게 되면 그 가능성이 정말 실현될 수 있다고 하더라도 우리가 그 길을 선택할 이유가 있는지 묻지 않을 수 없다. 앞서 말한 무엇을 개발할 것인가를 선별하는 두 번째 기준에 관한 질문이다. 즉 그것이 인간에게 도움이 되는 일인가라는 질문이다. 그 질문에 긍정적인 답을 얻기 전에는 누구도 - 적어도 자진해서는- 그 개발 기획에 매진하지는 않을 것이다. 나는 인간이 그 질문에 그리 쉽게 긍정적인 답을 내리지는 않을 것이라고 생각한다. 탈육화의 완성태라고 할 수 있는 그 상상의 내용은 사실상 인간 종의 진화가 종점에 도달한 상태를 그려본 것이기도 하다. 인간 종의 진화는 다른 모든 생물체와 마찬가지로 몸을 통하여 진행되어 온 것이다. 몸의 알고리즘이 몸을 떠나면, 자연은 진화를 주관하는 권리를 더 이상 행사하지 못한다. 이제부터 알고리즘의 업그레이드가 진화를 대신하는 것이다. 업그레이드의 주체는 일단 인간이겠지만, 곧 AI 자신이 될 것이다. 인간에게는 더 이상 자신을 변화시키기 위해 해야 할 일이 주어지지 않게 되는 것이다. 할 일이 더 이상 주어지지 않는 그런 상태를 천국의 상태로 생각할 사람들도 없지는 않을 것이다. 실제로 인간은 생존을 위해 그리고 더 나은 삶을 위해 갖가지 고생을 겪으며 살아왔다. 진화의 역사가 그러한 것이고 인간이 진화를 보조하기 위해 사용한 기술이 바로 그런 고생의 역사이기도 하다. 이제 AI가 모든

것을 맡아주는 것이 천국의 구원이 아닐 수 없겠다. 하지만 그런 천국의 구원을 우리는 진정 원할까? 몸을 지닌 인간 개체에게 그런 구원은 곧 죽음을 의미하는 것이다. 인간 종의 멸종을 의미하기도 한다. AI가 꾸며줄 미래를 포스트 휴먼의 시대라고 진단하는 것은 적어도 현재의 관점에서는 죽음의 선고로 들릴 수 있다. 인간이 멀쩡한 정신을 가지고 그 길을 선택할 리는 없을 것 같다. 인간은 AI 기술 개발을 통해 몸을 가지고 사는 삶이 계속 풍요롭게 되기를 기대하는 것이지, 일거에 할 일을 다 빼앗기고 죽음에 이르기를 원하지 않는다. 인간의 뇌도 계속 진화하면서 뇌의 기능을 전부 망라하려는 알고리즘 개발의 맹목적 야심을 앞질러 갈 것이다. AI가 미래의 어느 때 신처럼 출현해서 우리의 삶을 통째로 빼앗아가지는 않을 것이다.

그런데 이 경우 인간이 전적으로 자신의 의지에 따라 자신의 미래를 선택할 수 있는지 의심해볼 수 있다. 인간이 할 수 있는 일 중 맡기고 싶지 않은 일을 해내는 AI, 아니 아예 인간이 할 수 있는 일 모두를 다 맡아서 해낼 AI의 출현을 막고자 하면 과연 인간이 그렇게 할 수 있을지 얼른 자신 있게 말하기 어렵다. 우리는 지금까지는 AI 개발이 인간의 선택에 따라 이루어졌다고 믿고 있다. 그러나 그 역시 조금 더 깊이 들여다보면 단적으로 인간의 선택이라는 이야기를 하기 망설여진다. 도대체 기술 발달의 역사는 그 거시적인 흐름에 있어서는 필연의 경로를 밟는 것 같아 보인다. 다만 비교적 작은 규모의 시간 단위로 흐름을 쪼개어 보면 인간의 선택이 어느 정도 작용했다는 판단을 할 여지가 발견된다. 알파고는 개발자들의 선택이 있었기에 출현할 수 있었지만, 그 선택은 필연의 틀 안에서 이루어진 것으로 보일 수 있다. 알파고의 개발자들은 분명히 언젠가 함께 모여 바둑을 개발목표로 공략한다는 결정을 했고, 그 결정에 따라 프로그램 개발이 시작되어 소기의 성과를 올리게 된 것은 틀림없는 사실이다. 그러나 바둑이나 체스는 지적인 게임이고 세계적으로 많은 동호인들이 있다는 사정이 언제가 되었든 AI 개발자들의 관심을 끌게 되리라는 것은 미리 정해진 것이나 다름없다고 할 수 있다. 앞으로 우리들의 의사와 전혀 상관없이 정해진 필연의 진행

경로를 밟아 AI 개발이 갈 데까지 갈 것이라고 단언하는 것은 성급한 예단일지 모른다. 그렇지만 그것이 인간의 의지에 따라 전적으로 선택되는 것이라고 하기도 어답는 것이 진상일 것이다. 문제는 앞으로 AI 개발의 진행 과정에서 인간의 선택이 작용하는 폭이 계속 줄어들 가능성이 있다는 것이다. AI 기술이 제공해주는 편익에 익숙해지면 익숙해질수록 AI 기술에 의존하는 정도가 높아질 것이고, 사람들은 점차 AI 기술이 제공하는 서비스에 관한 비판적인 자세를 취할 수 없게 된다는 전망이 전혀 근거가 없는 것이 아니다. 일단 사람들이 그렇게 길들여지면 AI는 사람들의 간섭을 받지 않고 마치 스스로 자신의 앞날을 결정할 수 있는 힘을 가진 것처럼 발전해 나가게 될 것이다.

다른 기술과 비교해 본다면 AI 기술이 인간 개입을 배제하는 발전 시나리오가 현실이 될 가능성이 특별히 높다. 인간처럼 자유의지를 가지고 선택을 할 수 있는 강한 인공지능의 수준에 도달하지 않더라도, AI는 그 전 단계에서 학습능력을 획득해 가질 수 있다. 학습능력의 핵심은 자신의 앎을 스스로 개선할 수 있다는 데에 있다. 이런 능력을 갖춘 AI는 인간을 배움의 모델로 삼아 스스로를 개선해 나가는 과정에서 인간이 자신의 일을 맡기기 위해 컴퓨터 프로그램을 제작해내는 그 전략을 배울 수 있다. 게다가 AI가 인간이 하는 일을 외주받아 수행하면, 인간보다 훨씬 뛰어나게 할 수 있다. 그것은 이미 확인된 사실이다. 알파고는 바둑을 배운 뒤 그야말로 순식간에 세계 최강의 실력을 보유한 고수가 되었다. 다른 분야에서도 인간을 능가하는 AI의 능력은 전문가가 아닌 일반인에게도 잘 알려져 있다. 그렇다면 미래의 AI가 자신이 할 일을 맡길 또 다른 AI, 즉 AI의 AI에 해당하는 프로그램을 제작해내는 능력에 있어서도 인간을 능가할 가능성이 크다. 그 결과로 출현한 새로운 AI는 자신을 제작한 AI의 지능을 능가하는 것이 될 것이고 따라서 애당초 AI를 제작한 인간의 지능은 훨씬 더 격차를 두고 능가할 것이다. 이런 프로그램이 등장하는 때를 기점으로 지능이 가히 폭발적이라 할 수 있는 방식으로 증강될 것은 필연적인 일이다. 인간은 개입하지 못하고 그냥

지켜보는 사이에 AI는 재귀적recursive 자기 개선의 필연적 경로를 밟아 지능 폭발을 거쳐 초지능超知能superintelligence이 되는 것이다.

이와 같은 시나리오가 이야기된 것은 반세기 전의 일이다. 그러나 지능 폭발의 귀결인 초지능에 관한 자세한 이야기는 새로운 세기에 들어설 즈음이 되어서야 사람들의 관심을 끌게 된다.[5] 사람들은 AI개발의 성과를 목도하고 나서야 초지능의 가능성도 심각하게 생각하기 시작했고 초지능의 출현이 초래할 변화에 대한 자세한 이야기에도 관심을 갖게 되었다. 그 이야기의 핵심은 지금까지 자신을 만물의 영장, 즉 모든 생명체 중 가장 지능이 높은 존재로 알고 지내왔던 인간이 처음으로 자신보다 훨씬 지능이 높은 존재와 같은 세상을 살게 되는 상황을 예상하는 것이다.

인간으로서는 처음 겪게 될 이 상황에 대한 전문가들의 예상은 낙관론과 비관론으로 서로 갈려 있다. 어느 쪽이 더 설득력이 있는지 비전문가의 입장에서 판단한다면 자연스럽게 지금까지 인간이 자신보다 지능이 낮은 생명체들을 어떻게 취급하며 살아왔는지를 먼저 생각해보게 된다. 인간이 한 짓에서부터 유추하자면 일단은 비관론 쪽으로 기우는 것은 어쩔 수 없다. 초지능이라고 해서 인간의 척도에 맞는 도덕성을 갖추어야 하는 것은 아니기 때문에 초지능을 갖춘 존재가 인간을 도덕적으로 훌륭하게 대접할 것이라는 보장은 없다. 다른 생명체에게 끔찍한 짓을 자행해왔던 인간은 그나마 어느 정도라도 도덕성을 갖추고 있는 존재다. 하지만 아예 도덕이 무엇인지 모르는 초지능의 행태는 어떨까? 인간이 다른 생명체에게 한 짓보다 훨씬 더 끔찍한 짓을 인간에게 할 수 있을 것이다. 물론 그런 경우 끔찍하다는 것은 인간의 입장에서 하는 이야기다. 초지능의 입장에서는 - 도대체 그런 입장이라는 것이 있다면 - 그 어떤 도덕 판단의 대상이 될 행태가 아니다.

---

5) 이 시나리오는 튜링의 동료였던 수학자 굿 I.J. Good (1965)이 이야기한 것이다. 굿의 이야기는 마침 본격적인 AI연구가 시작되었던 때에 나왔다. 당시 분위기에서 그가 이야기하는 시나리오에 시사된 미래의 문제는 심각하게 논의될 수 없었다

그 점은 여러 가지 방식으로 부연할 수 있겠는데, 그 중 소위 '도구적 수렴instrumental convergence'의 개념에 의거한 시나리오의 부연이 특히 그럴 듯하다. 도구적 수렴이란 어떤 종류의 목적이든 - 그것이 아무리 시시한 목적이든 상관없이 - 그것을 달성하기 위해서는 일단 확보해 놓고 보아야 할 기본적인 도구적 수단이 있다는 것이다. 초지능의 경우에 목적하는 바가 가령 은하계에 존재하는 모든 모래알의 총 개수를 세는 일이라고 하면 그 목적을 달성하기 위해서 초지능은 우선적으로 그 일을 해낼 자신의 존재를 계속 보존해야 한다. (자기 존재의 보존은 이 경우 목적이 아니라 수단으로서 가치를 지닌 것이 된다.) 그 다음으로 언제 끝날지 알 수 없는 그 일을 수행하기 위해 필요한 동력 자원을 충분히 확보하는 것이다. 그러니까 문제의 초지능은 자신의 존재를 위협할 수 있는 요인을 완전히 제거하고 나아가 가능한 한 동력 자원 공급원을 자신이 쓸 수 있는 소재로 바꾸어 확보해 놓으려 들 것이다. 그렇다면 초지능에게는 그 다음에 할 일 중 하나가 곧 인간을 멸살하는 것이 될 것이다. 초지능의 입장에서 보면 인간의 존재가 자신의 존재를 위협하는 요인 중 하나이기 때문이다. 인간은 초지능이 모래알을 세는 무의미한 일을 하기 위해서 동력 자원을 무한히 확보하려 드는 것이 자신들의 존재에 대한 위협이 되니 초지능을 제거하려 들 것은 당연하다. 뿐 아니라 초지능은 인간의 육신이 곧 자기가 확보하고자 하는 동력 자원 공급원이 된다는 것을 인식할 터이니, 동력을 공급하는 소재로 인간을 가공하려 들 것이다. 이런 짓을 하려는 초지능을 모자라는 지능을 지닌 인간이 과연 막을 길이 있을까? 미래의 초지능이 어떤 목적을 추구하든 어떤 과제를 수행하든 - 인간이 주문하는 일이든 자신이 어떤 이유로든 결정해서 하려는 일이든 - 그것은 인간 세상을 생지옥으로 만들거나 아예 인간을 멸종에 이르게 하는 일로 귀결될 가능성을 배제할 수 없다.[6]

---

6) 수단적 수렴의 개념은 대학의 교재로 많이 쓰인 S. Russell과 P. Norvig (2003)의 저술을 통해 널리 알려졌는데, N. Bostrom (2014)은 그 개념을 PI의 소수점 이하 자릿수를 계속 계산하거나 종이집게를 만들어 내는 일, 모래알을 세는 것처

그에 반해 초지능이 인간에게 천국의 삶을 가능하게 해줄 수 있다는 낙관론은 사실 설득력이 별로 높지 않다. 불치병이 다 치료되고, 아주 오래 살 수 있고, 원한다면 영원히 죽지 않을 수도 있다는 것이 그 그림의 구체적인 내용 전부다. 나머지는 좀 막연한 내용이어서 얼마나 좋은 것인지 정확히 알 수도 없다. 죽으면 천당에 가서 영원히 행복하게 산다는 세속 신앙의 그림을 더 자세히 그려서 그 행복이 무엇인지 알게 해달라고 주문한들 누구도 그 주문에 맞는 그림을 그려 보일 도리가 없을 것이다. 인간은 지옥을 상상하는 데에는 재주가 있지만 천당을 상상하는 데에는 무력하다. 그런데 천국의 삶이 구체적으로 어떤 것이든 간에, 낙관론을 주장하는 쪽은 초지성이 인간에게 그런 삶을 마련해줄 것이라고 예상할 근거가 무엇인지도 이야기해야 한다. 이 대목에서 낙관론은 내놓을 이야기가 충분치 않다. 지능이 높으면 인간에게 좋은 일을 할 마음이 저절로 생길 것이라는 정도의 이야기를 한다면 그것은 그저 천진한 희망사항을 피력하는 것일 뿐이다.

초지능이 인간에게 비도덕적인 일을 하지 않도록 하게 하는 길은 오직 너무 늦기 전, 그러니까 인간의 개입이 가능한 시점에서부터 초지능이 도덕적 사고를 할 수 있게끔 개발의 방향을 잡는 것이겠다. 최근 도덕적 사고를 하는 인공적 행위자Artificial Moral Agent에 관한 논의가 연구자들 일각에서 활발해지고 있는데,[7] 거기에는 계속 강력해지는 AI의 힘이 의식되고 있다는 배경이 있을 것이다. 그런데 그 논의는 쉽게 끝날 것 같지 않아 보인다.

럼 특별히 하찮아 보이는 일을 예로 들어 설명을 했다. 그렇게 한 의도는 목적을 설정하는 지성의 수준과 목적을 달성하기 위해 수단을 동원하는 지능의 수준이 상호 독립적이라는 자신의 명제orthogonality thesis와 수단적 수렴의 개념을 연결하여 초지능이 하는 일의 위험성을 좀 더 명확하게 설명하려는 것이었다.
M. Tegmark (2018)는 프로메테우스라는 가상의 초지능을 주역으로 하는 좀 더 현실감이 있는 자세한 스토리를 구상해보기도 했다. 그러나 실제로 초지능의 목적 설정에 관련된 능력이 어떤 것인지, 초지능이 출현하면 그것이 도대체 어떤 목적을 추구하게 될지는 정말 알 수 없는 일이다

7) 도덕적인 AI의 제작 가능성 및 제작 전략 그리고 연구 현황에 관해서는
W. Wallach & C. Allen (2009)를 참조할 수 있음.

윤리학적 논의의 긴 역사를 잠깐 머리에 떠올리기만 해도 그 논의는 도덕성이 무엇인지, 구체적인 행위 패턴에 대한 윤리적 판단을 어떻게 할 것인지 등 철학적인 문제를 놓고 계속 더 시간을 쓸 것이라는 짐작을 할 수 있다. 아마 늦기 전이 아니라 늦은 것이 확실한 시점을 지나서까지도 그 논의는 계속될 가능성이 크다.

그렇다면 초지능이 도덕성을 갖게끔 조처하는 것보다는 초지능이 도덕적으로 옳고 그르고 간에 덮어놓고 인간을 사랑하는 마음을 갖도록 하는 것이 더 간명한 방안일 것이다. 실제로 인간에게 호의를 가진 인공지능Friendly Artificial Intelligence의 개발 가능성을 타진하는 연구자들도 있다. 그야말로 첨단 중의 첨단 연구라 할 수 있겠는데, 사실은 그런 연구의 착상 계기는 종교를 통해 우리에게 이미 잘 알려진 것이다. 전지, 전능한 신의 존재를 상정한 종교는 그 신이 인간을 무조건 사랑한다는 상정까지 하고 있다. 그런 상정을 하지 않으면 종교는 인간에게 그 어떤 구원의 희망도 줄 수 없을 것이기 때문이다. 인간에게 호의를 갖는 초지성의 착상도 마찬가지다. 그 착상은 인간을 사랑하는 신에게서 인간 구원의 최종적 보장을 받고 싶다는 희망의 컴퓨터공학적 버전이라 할 수 있는 것이다. 과연 그와 같은 초지능이 출현할 수 있을지 강한 인공지능과 마찬가지로 확실하게 알 길은 없다. 그러나 인간에게 호의적인 AI의 출현을 바라는 것은 성경 속의 신이 현실의 세계에 강림하는 것을 바라는 것과 비슷해 보인다. 그것은 인류의 구원을 위해 컴퓨터공학이 Deus ex machina의 역할을 맡아주기 희망하는 것이 아닐까 한다.

이제 거의 종교적인 수준의 희망까지 언급하였으니까 이야기를 마무리해야 할 때가 된 것 같다. 우리는 AI기술 개발이 지향하는 바가 기술 발달의 역사를 관통해온 탈육화의 최대한을 달성하는 것이라는 사실을 확인했다. 그러면서 동시에 뇌를 사용하는 수고까지 포함해 모든 육신의 수고에서 완전히 벗어나고자, 인간이 할 일을 전부 AI에게 맡기고 진화의 여정을 끝내는 것은 인간의 멸종을 뜻한다는 생각도 하게 되었다. 우리가 진정 바라는 바는 간단치 않은 것이다. 우리는 AI의 권한을 미리 제한하여 인간 삶의

도우미로 이용할 수 있기를 바란다. 그러니까 AI개발에 매진하면서도 맡길 일과 맡기지 않을 일은 가려내면서 개발 방향을 결정하고자 하는 것이다. 그러나 적극적으로 그런 선택을 하기 전에 개발 도상에서 초지능이 출현하는 사태를 겪을 수 있다는 가능성의 문제에 봉착했다. 그것은 AI 개발 노력이 본래 지향한 바를 향해 나아가는 중 돌연히 출현한 것이어서, 그러한 초지능이 인간의 지능을 모사한 존재, 다시말해 인간을 닮은 존재인지도 확실하지 않은 존재다. 그런 존재를 인간 삶의 도우미로 이용할 수 있으리라는 보장은 전혀 없다. 그보다는 그것이 인간 삶을 생지옥으로 만들거나 아예 인간 종을 멸살할 가능성이 더 커보인다. 지금부터 그런 위험에 대비하고자 한다면 AI개발의 우선 순위를 조정해서 다른 무엇보다도 도덕적 사고를 하거나 인간을 무조건 사랑할 줄 아는 AI의 개발 기획을 최우선으로 추진해야 할 것처럼 보인다. 그러나 우리가 그런 일을 해낼 수 있을지 회의하지 않을 수 없다. 인간이 성경 속의 신의 존재는 상상할 수 있지만, 그 신을 제작해서 현실 속에 존재하게끔 하는 일까지 해낼 능력까지 가졌을 것 같지는 않기 때문이다. 진화의 여정에 있는 인간은 그런 일을 희망할 수 있는 정도의 지능을 갖게 된 지점에 도달한 것은 틀림없다. 그러나 그것이 인간에게는 진화의 여정에서 도달한 최종의 지점일 수 있다.

참고문헌

Berkeley, I. & C. Rice. (2018), “Machine Mentality?” in Vincent C. Müller (ed.), *Philosophy and Theory of Artificial Intelligence*, pp. 1-16, Berlin: Springer.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford: OUP.

Good, I.J. (1965), “Speculations Concerning the First Ultraintelligent Machine”, in Franz L. Alt and Morris Rubinoff (eds.), *Advances in Computers*, pp. 31–88. New York: Academic Press.

Hayles, N. K. (1999), *How We Became Posthuman,* Chicago: The University of Chicago Press.

Kurzweil, R. (2005), *Singularity is Near: When Humans Transcend Biology*, London: Penguin.

Russell, S. & Norvig, P. (2003), *Artificial Intelligence: A Modern Approach*, N.J: Prentice Hall.

Scheutz, M. (2014) “Artificial Emotions and Machine Consciousness” in K. Frankish & W. Ramsey, (eds.), *Cambridge Handbook of Artificial Intelligence*, pp 247-354, Cambridge: Cambridge University Press.

Searle, J. (1980), "Minds, brains and programs", *Behavioral and Brain Sciences* 3: 417–24.

Turing, A. (1950), “Computing Machinery and Intelligence”, *Mind,* 59: 433-60.

Wallach, W. & Allen, C. (2009), *Moral Machines*, Oxford: OUP.

“Oldowan-tradition stone chopper”, https://en.wikipedia.org/commons/oldowan. (검색일: 2019.03.02.)

Abstract

Achieving whole brain emulation by closely modelling the computational structure of cerebral operation is the final goal of AI research. We can saflely assume that the success in such ambitious software engineering will make the advent of so-called Strong AI possible. Now, - setting the problem of feasiblity aside - we have to ask a question to what extent Strong AI would serve useful purpose. However, more importantly, we have to reckon with the possibility that, before we can answer the question concernig the Strong AI, we might encounter with another far more urgent problem; even on the level of Weak AI a sort of Superinteligence can possibly begin its existence. In this paper I will discuss the possible advent of Superintelligence and make some reflections on what we can do to safeguard against the dangers which Superintelligence might incur.

【Keywords】 Technology Development, Outsourcing of Bodily Functions, Algorithm, Strong AI, Evolution

논문 투고일: 2019. 03. 28
심사 완료일: 2019. 04. 15
게재 확정일: 2019. 04. 15