인공지능의 철학적 문제

자오팅양(趙汀陽) / 중국사회과학원 철학연구소 연구원

기술에 의해 촉발된 새로운 존재론적 문제

기술 문제가 존재론적 문제가 될 때 진정 중요한 가능성이 드러난다.

존재론은 과거로부터 지금까지 단일 주체의 인식론적 시야(horizon)의 제한을 받아왔다. 즉, 인간의 시야로 존재를 사고해왔고 인간의 시야를 유일한 주체적 시야로 묵인해왔다. 따라서 지금까지 존재론은 인식론을 초월한 적이 없다. 칸트는 바로 이러한 인간의 인식론적 자신감에 근거하여 인간이 자연의 입법자라고 주장하기까지 했다. 인간이 자연의 입법자이므로 인간의 시야가 유일한 시야라는 순환논증도 가능하다. 그러나 사람들은 신이 무한한 가능 세계를 볼 수 있다는 라이프니츠의 논증처럼 신학적으로 인간보다 더 높은 절대 시야를 상상했다. 그러나 이러한 이론적 절대 시야는 인간의 몫이 아니며 인간이 무한한 가능 세계의 절대 시야가 어떤 모습일지 상상하는 것은 불가능하다. 인간에게 허락된 유일한 시야는 인간의 주체적 시야이고 비트겐슈타인의 말을 빌리자면 이것이 사고의 한계이다.

눈이 눈 자체를 볼 수 없는 것처럼(비트겐슈타인의 비유) 생각은 스스로를 초월할 수 없다. 하지만 생각으로는 불가능한 일이 현실에서 실현될 수도 있다. 인공지능은 또다른 주체, 또 다른 입법자 또는 또다른 눈으로 성장할 가능성이 있다. 이는 일방향의 존재론이 쌍방향(심지어 다방향)의 존재론으로 변화할 수 있다는 존재론적 차원의 거대한 변화를 의미한다. 그렇게 되면 세상은 하나의 주체의 시야에 속할 뿐 아니라 두 개 이상의 주체, 심지어 인간 이외의 새로운 주체에 속하게 될 것이다. 인공지능이 새로운 주체가 될 수 있다면 세상은 새로운 존재론으로 진입하게 될 것이다.

인공지능에 대한 정의는 다양하다. 과학에서는 통상 튜링기계 개념에 속하는 인공지능을 AI라고 하고, 인간 지능에 해당하는 인공지능은 AGI(범용인공지능, artificial…general…intelligence), 인간 지능을 전면적으로 능가하는 고급 지능은 SI(초지능, super…intelligence)라고 한다. 이러한 과학적 분류는 기술적으로 측정 가능한 지능의 등급을 보여주지만 우리는 지능의 철학적 성격, 즉 그것이 'Cogito'의 주체성을 갖추었는지를 논하고자 한다. 따라서 본고에서는 인공지능을 철학적 성격에 따라 분류할 것인데, 그 중 하나가 AI라고 불리는 것이다. 이는 데카르트의 'Cogito' 기준에 도달하지 못한 비성찰적(non-reflexive) 인공지능으로, 그 적용 범위는 과학에서 분류하는 AI와 대체로 일치하며 튜링기계(AlphaGo 같은 단일 기능 인공지능 및 아직 성공하지 못한 복합 기능 인공지능 포함)의 개념에 속한다. 다른 하나는 ARI라는 것으로, 데카르트의 'Cogito' 기준을 넘어서는 성찰적 인공지능(artificial…reflexive…intelligence)이다.

ARI는 초인공지능 또는 하이퍼 튜링기계와 거의 비슷한데, 필자는 그것이 가지고 있는 자체 시스템 성찰 능력을 드러내기 위해 '괴델기계'라고도 부른다. 여기서 ARI는 AGI 또는 SI를 반드시 포함하지만 ARI가 반드시 AGI 또는 SI가 되는 것은 아니라는 점에 주의해야 한다. 이는 ARI가 인간의 모든 재능을 가지고 있지 않을 수 있지만 자기 성찰 능력 및 자체 시스템 수정 능력을 가지고 있어 자율적 주체성을 구비함으로써 지배 불가능한 타자의 마음이 되고 세상의 또 다른 주체가 된다는 것을 의미한다.

주체성을 기준으로 한 분류는 인공 지능의 가능한 질적 변화, 즉 특이점을 부각시키려 한다. 현재로서는 인공지능에 질적 변화가 발생하는 특이점은 아직 요원하고 예언자들이 그 가능성을 다소 과장하는 것처럼 보이기는 하지만, 문제는 인공지능의 특이점이 가능하다는 것이다. 지능의 핵심은 연산 능력이 아니라 성찰 능력에 있다. 인간 주체성의 본질은 성찰 능력에 있으며, 그것이 없이는 사고의 주체가 될 수 없다. 인공지능에 성찰 능력이 없다면, 연산 능력이 강할수록 인간에게는 더 유용하고 치명적인 위험은 없을 것이다. 예를 들어 AlphaGo…Zero는 강력한 연산 능력을 가지고 있지만 인간에게 위협이 되지는 않는다. 그 반대로 인공지능이 다른 모든 면에서 인간보다 약하더라도 성찰 능력을 가지고 있다면 위험한 주체가 된다. 만약 인간에게 가능한 대부분의 기능이 결여되어 식량이나 석유 같은 것은 생산하지 못하면서 첨단 무기를 제조해서 사용할 줄 알고 자기 성찰 능력을 강화하는 인공지능이 있다면 그 결과가 어떨지 가히 상상이 된다. 지금까지의 인공지능은 알고리즘 능력이나 뇌와 유사한 신경 반응 능력만 있을 뿐 성찰 기능은 갖추지 못했고 성찰 기능을 갖출 수 있을지 여부도 단언하기 어려우므로 여전히 안전한 기계라 할 수 있다. 설사 미래에 유연한 반응능력을 갖춘 다목적 인공지능이 출현한다 하더라도 성찰 기능이 결여된다면 새로운 주체가 되지 못한 채 인간의 가장 강력한 어시스턴트로서 머물게 될 것이다.

대부분의 기술은 인간의 능력을 강화하거나 확대할 뿐이다. 예를 들어 생산 및 제조 수단인 기계는 증기기관에서 발전기로, 자동차와 항공기에서 우주선으로, 그리고 전화와 컴퓨터, 인터넷에서 양자기술 등으로 강화 및 확대되었다. 튜링기계 인공지능도 이 범주에 속한다. 기술이 아무리 강력하더라도 기술 시스템 자체에 성찰 능력이 없는 한 존재론적 차원의 위험은 없다. 낙관적으로 볼 때 이러한 기술로 인한 사회적, 문화적 또는 정치적 문제는 여전히 인간이 통제 가능한 범위 내에 있다. 물론 일부 고위험 기술과 심지어 악성 기술이 존재하기도 한다. 예를 들어 고위험 시설인 원자력발전소에서 핵폐기물을 처리할 확실하게 안전한 방법이 아직은 없고, 핵무기의 주요 기능은 대량 살상이다. 인공지능과 유전공학기술의 발달은 능력의 향상이라는 개념을 넘어 종을 변화시키거나 새로운 종을 만들어내는 기술로 변모하고 있어 인간이 감당할 수 없는 위험을 내포하고 있다. '신과 같은' 능력을 가진다는 것은 아직은 이론적 가능성에 불과하지만, 실천에 앞서 새로운 존재론적 문제가 제기되었고 그와 연관되어 새로운 인식론적, 정치적 문제도 제기되었다. 여기서 주의할 점은 새롭게 대두된 철학적 문제가 전통적인 철학적 문제의 업그레이드된 버전이 아니라 지금까지 없었던 전혀 새로운 문제라는 것이다. 따라서 장자, 베버, 헤겔과 같은 전통 철학의 기술에 대한 비판은 이들 새로운 문제들에 있어서는 전혀 유효하지 않으며 아무런 관련성도 없다. 즉, 인본주의적 윤리관 또는 가치관은 기술 분야의 새로운 문제에 답하기 어렵다는 것이다.

인공지능과 유전공학기술 모두 인간 개념에 도전하는 존재론적 문제를 제기하지만 인공지능의 위험성

이 유전공학기술보다 더 큰 것 같다. 인공지능이 진정한 창조물이어서 완전히 예측 불가한 반면 유전공학기술은 종의 개량에 연관되므로 자연적인 한계가 있기 때문이다. 이러한 주장은 증명하기는 어렵지만 사실일 수 있는 신념에 기반을 두고 있다. 즉, 하나의 폐쇄적 시스템에서 내부 요소의 혁명 능력이 전체적으로 정해진 물리적, 또는 생물학적 한계를 넘어설 수 없으며 내부 변화가 전체 한도를 초과하는 경우 시스템이 붕괴된다는 것이다. 이는 유전공학기술의 혁명성이 생명의 생물학적 한도를 넘어설 수 없다는 것을 의미한다. 유전공학기술이 종을 최적화하는 방법이 될 수 있지만 어떠한 종이든 생명으로서 허용되는 변화의 한도가 있기 마련이다. 유전공학기술이 진정 인간의 불로장생을 가능하게 할 수 있을지는 여전히 미지수이다. 일부 파충류나 어류는 성장속도가 느려서 수명이 길고 작은보호탑해파리는 어린 개체로 돌아가는 특이한 기능으로 인해 영생불사하는 것 같다고 하지만 실망스럽게도 수명이 매우 긴 생물은 지능이 극히 낮다. 만약 인간 유전자를 근본적으로 개조한다면 생명시스템이 붕괴되지 않을까? 예를 들어 뇌나 면역체계가 붕괴되는 건 아닐까? 유전공학기술을 통해 인간을 신과 같은 완전히 새로운 종으로 개조하는 것은 생물학적으로 불가능해 보인다. 더욱 현실적인 문제는 유전자 최적화는 설사 그것이 제한적이라 하더라도 사회 문제를 악화시켜 인류의 재앙을 초래할 가능성이 매우 높다는 점이다.

새로운 종을 창조하는 능력 면에서 인공지능은 유전공학기술보다 더 위험하다. 인공지능이 특이점을 돌파한다면 예측 불가한 새로운 주체가 만들어질 것이고 새로운 주체에게 있어 기존의 일원적 주체의 지식, 시야 및 가치관은 붕괴될 것이며 이원적 주체(심지어 다원적 주체)의 세상은 아직 상상하기 힘들 것이다. 많은 SF작품들이 가공할만한 로봇이나 외계인을 그려냄으로써 사람들에게 피학적 쾌감을 가져다 주기는 했지만 인간은 기술이 지배하는 미래에 대해 제대로 된 사상적, 심리적 준비를 하지 못했다. 아직 요원한 이원적 주체의 세상은 차치하고 사람들은 눈앞의 초보적 인공지능화 또는 유전공학화된 사회에 대해서조차 충분한 경각심을 가지지 못하고 있다. 종말 문제를 생각하지 않더라도 고도로 기술화된 사회에서도 기존의 문제점이 극대화되어 인류가 빈부격차, 계급투쟁, 인종간 투쟁, 민족간 투쟁, 자원부족, 자연의 쇠퇴와 불균형과 같은 해결 불가한 딜레마에 빠질 것이다.

류츠신(刘慈欣)은 자신의 에세이 형식의 단편소설 <Taking Care of God>(贍養上帝), <The Wages of Humanity>(贍養人類)에서 모든 것이 스마트화된 "노년 문명"의 절망적인 이야기를 그리면서 두 가지 핵심적인 주장을 했다. 그 첫째는 고도로 발달한 인공지능이 거의 전능하고 전자동으로 운영되므로 전 인류가 풍요로운 삶을 누리게 할 수 있는 "기계 요람"을 만들어준다는 것이다. 이와 관련하여 작품 속 우주에서 고도로 발달하여 "신의 문명"이라 불리는 사람은 이렇게 말한다. "스마트한 기계는 물질적이건 정신적이건 우리가 필요로 하는 모든 것을 제공할 수 있으므로 우리는 생존을 위해 어떤 노력도 할 필요가 없이 안락한 요람에 누워 있는 것처럼 온전히 기계에 의지해서 살아갈 수 있어요. 한번 생각해보세요. 지구의 숲에 아무리 따도 다하지 않을 과일이 가득하고 손만 뻗으면 잡을 수 있는 사냥감이 사방에 널려 있었다면 유인원이 인간으로 진화할 수 있었을까요? 우리는 기계 요람이 만들어준 풍요의 숲에서 점차 기술과 과학을 망각했고 문화는 산만하고 공허해졌으며 혁신능력과 진취성을 상실했고 문명의 노화가 가속화 되었다"라고 하고, 따라서 모든 인류가 "1원2차 방정식조차 할 줄 모르는" 쓸모 없는 존재가 되었다고 지적했다(Taking Care of God). 여기서 제기되는 문제는 인공지능이 만들어낸 행복한 삶이 인류의 바람과 달

리 문명의 쇠퇴를 초래했다는 것이다. 둘째는 인공지능 사회에 보다 현실적인 또 다른 버전이 있다는 것이다. “신의 문명"만큼 발달하지는 못한 지능화된 문명은 이미 절망적인 상태에 빠졌다. 소설에서 지구보다 발달하고 유형적으로 완전히 유사하여 “지구의 형제 문명”에 속한 사람이 지구 문명의 앞날에 대해 이렇게 말한다. “완전히 지능화된 사회에서 노동이 필요 없게 되면 부자도 더 이상 가난한 사람을 필요로 하지 않게 되어 계층 이동의 사다리도 끊어지게 됩니다. 부자가 교육을 독점하게 되기 때문입니다. 그러한 교육은 전통적 의미의 교육이 아니라 인공지능과 생명공학의 합작으로 이루어진 인간-기계 통합기술로서, 이렇게 극도로 값비싼 ‘교육’을 구매하게 되면 슈퍼맨이 되어 모든 능력 면에서 전통적 인간과 완전히 다른 등급이 되며 그 등급 차는 인간과 동물의 차이보다 큽니다. 따라서 부자와 빈자는 가난한 사람과 개가 동일 종이 아닌 것처럼 서로 다르며, 가난한 자는 더 이상 인간이 아닙니다.....가난한 자에 대한 동정에서 핵심은 ‘동(同)’자에 있습니다. 쌍방이 동일 종으로서의 토대를 가지고 있지 않다면 동정 또한 있을 수 없습니다”라고(The Wages of Humanity)……. 이는 초인공지능이 출현하기를 기다릴 것 없이 지능화 사회만 되더라도 인류를 새로운 종으로 변화시킬 수 있다는 것을 말해준다. 다시 말해 인공지능의 특이점이 나타나지 않더라도 인류 문명에 심각한 문제가 도래할 수 있다는 것이다.

인공지능의 특이점은 잠시 논외로 하고 지금 우리에게 닥친 문제만 보더라도 충분히 놀랄 만하다. 인간은 원래 이익 배분, 사회 갈등, 집단 투쟁 또는 문명 충돌 등의 문제를 잘 해결하지 못해왔고, 이들 문제가 해결될 수 없었던 근본 원인은 인간 노력의 한계가 아니라 인간 본성의 한계에 있다. 실망스러운 점은 인류의 문제 해결 능력이 문제를 만들어내는 능력보다 훨씬 약해서 오래된 문제를 해결하기 어렵고 인공지능 또는 유전공학기술은 Amplifier 또는 Accelerator가 되어 오래된 문제를 더욱 악화시킨다는 것이다. 인류가 정치제도, 법제도, 윤리 체계의 발명과 같은 위대한 업적을 이루었지만 인간의 사고 능력은 한계에 가까워지고 있는 것 같다. 지난 수십년 간 전세계적으로 사상적 피로나 나태의 징후가 갈수록 분명해졌다. 창의적 아이디어가 크게 감소했고 사상적 틀과 개념은 200년 전 수준에 머물렀다. 인공지능 및 유전공학기술 등 새로운 문제에 대해서는 wishful thinking에 가까운 윤리적 비판 외에는 속수무책인 상태이다. 인공지능에 대한 윤리적 비판은 왜 이렇게 적절한 해답을 내놓지 못하는 것인가? 이와 관련하여 매우 우려할만한 상황이 있다. 그것은 바로 문명이 고도로 지능화된 세상에서 윤리학의 문제는 사라지거나 적어도 주변화될 가능성이 크다는 것이다. 이는 문명 발전에 대한 인류의 기대와 배치되면서 다소 황당해 보이지만 그 가능성이 매우 크다.

<Taking Care of God>에서 고도로 발달한 인공지능이 전 인류의 풍요로운 삶을 가능하게 할 수 있는 “기계 요람”을 형성했다.

문명의 인공지능화는 문명의 재야만화(re-barbarization)를 초래할 가능성이 높다.

인류 문명이 끊임 없이 '진보'하고 있다는 것이 보편적 믿음이다. 하지만, 과학기술이 진보하고 있다는 것은 의심의 여지가 없지만 과학과 기술을 제외한 다른 측면이 진보하는지 여부는 논쟁적이다. 기술의 본질은 능력이며 능력이 클수록 게임의 균형점이 기술을 가진 자에게 더 유리해진다. 소수가 기술을 장악하고 있다면 약자의 협상 이익이 작아질 것이다. 그렇다면 인간 본성이 변하지 않는 한 문명의 인공지능화는 문명의 재야만화(re-barbarization)를 초래할 가능성이 매우 높다. 여기서 '야만화'란 태곳적 생활수준으로 퇴보한다는 것이 아니라 사회관계가 힘이 진리가 되는 정글 상태로 악화되는 것, 즉 기술자원을 점유한 사람이 압도적인 필승 기술을 장악하고 있으면서 도덕, 법률, 정치가 필요 없게 되는 것을 의미한다. 사람들은 이러한 홉스적 진리를 잘 알고 있지만 이렇듯 유쾌하지 못한 문제를 회피하고 허망한 환상을 지키는 쪽을 더 선호할 뿐이다. 인류는 지금까지 운 좋게도 이러한 '최악의 세상'의 문제를 성공적으로 회피해왔다. 그 이유는 홉스적 세상에는 절대 강자가 없고 강자라고 하더라도 치명적 약점이 많아 모든 사람이 약자이기 때문인데, 모든 사람이 약자라는 사실이 인류에게는 행운이다. 니체가 지적했듯이 약자야말로 도덕을 필요로 한다. 모든 사람이 약자라는 점은 인류에게는 행운이자 도덕, 법률, 정치의 토대이기도 하며, 도덕, 법률, 정치는 서로에게 해를 끼칠 수 있는 약자들 간의 장기간에 걸친 게임을 통해 형성된 안정적 균형이다. 물론 게임의 균형으로 설명되지 않는 '정신이 물질에 우선하는' 예외가 있기도 하다. 예를 들어 이타적이거나 자기희생적 도덕 같은 것은 인간의 신비라고 할 수 있다. 하지만 정신이 물질에 우선하는 현상은 인간 사회의 주요 구조가 아니며 결정적인 변수도 아니다.

고도로 발달한 인공지능이나 유전공학기술은 언젠가 인간의 행운이 다했다고 선언할 수도 있다(필연적인 것은 아님). 최소 비용과 최대 이익이라는 논리에 따르면 인류 문명이 복잡한 제도와 윤리, 법률을 발전시킨 것은 저비용의 간단한 방식으로 힘과 이익 문제를 해결할 능력이 없었기 때문이라고 추론할 수 있다. 사람들은 일반적으로 문명의 복잡성이 문명의 발달 정도를 나타내며 복잡성과 정교함, 교묘함, 조화, 난이도, 정신 등 문명의 지표 간에 확실한 상관성이 있기 때문에 '고차원적'이라고 믿는다. 성숙한 문명의 윤리도덕, 법률과 제도, 사상과 예술은 모두 복잡하다. 문명의 성숙함을 나타내는 이들 지표는 한 가지 본질적인 문제를 은폐한다. 그것은 바로 복잡성은 고비용(거래비용 포함)을 의미하고, 고비용이기 때문에 이익을 극대화할 수 없다는 것이다. 그래서 한 가지 잔인한 논리가 있다. 최소 비용의 가장 단순한 방법으로 최대의 이익을 얻을 수 있는 능력이 있다면 사람들은 간단하고 거친 방법으로 문제를 해결하려고

하지, 복잡하고 고비용이 드는 방법을 선택하지 않을 것이라는 것이다. 따라서 인공지능과 유전공학기술이 절대적인 강자를 만들어낸다면 절대 강자는 절대 우위의 기술을 이용하여 문명의 재야만화를 이룰 가능성이 높다는 것을 알 수 있다. '쓸모 없는' 인간을 없애버리고 복잡하고 고비용이 드는 윤리, 법률, 정치를 포기하는 것이 그 예이다. 재야만화된 첨단기술의 세상에서 윤리학이 해답을 제시할 수 없을 것임은 자명하다. 그러나 인간에게는 아직 성찰하고 조정할 시간이 남아있다. 그런데 이것이 good news라고 할 수 있는지는 잘 모르겠다.

미래에 출현 가능한 문명의 재야만화에 대한 사람들의 경각심이 부족한 것은 아마도 계몽운동 이후 인간의 주체적 오만함과 관련이 있을 것인데, 이는 이성적 오만함이기도 하다. 계몽된 이성은 신에 대한 숭배에서 벗어나 인간을 숭상하기 시작했고, 이 위대한 사상혁명은 인간이 주체적 승리에 도취되어 인간으로서의 진면모를 점차 망각하게 했다. 신을 숭배하던 고대에는 신이 의심할 수 없는 대상이었고 인간을 숭상하게 된 현대에는 사람이 의심 불가의 대상이 되면서 인간의 약점, 결점 심지어 죄악까지도 엄폐되었다. 세상에 나쁜 일이 생기기만 하면 제도나 관념을 탓하고 인간 자체를 성찰하지 않는다. '원죄'에서 벗어난 인간은 아무런 부담도 거리낌도 없이 인간이라는 이름으로 모든 쾌락과 이익, 권리의 획득을 요구한다. 현대 정치의 근거는 더 이상 인간을 구속하는 자연신학이나 종교신학이 아니라 인간의 신학, 소위 숭고한 인간이다. 그러나 미약하고 이기적인 인간이 설사 '숭고하다'고 한들 얼마나 숭고할 수 있을까? 인간은 무엇에 기대어 원하는 모든 것을 얻는가? 주체적 오만함은 오히려 인간의 신학이 반인간적임을 보여준다. 개인은 보호할 수 있으되 인류는 보호할 수 없다는 개인주의의 치명적 약점은 전 인류가 도전에 직면했을 때 여지없이 드러난다. 미래에 초인류 인공지능이 등장하거나 극소수의 사람들이 초능력을 가진 인공지능을 통제하게 된다면 개인주의 사회는 인공지능의 통치에 저항할 수 없을 것이다. 왜냐하면 인공지능은 개인이 아니며, 모든 개인보다 훨씬 강력한 시스템이기 때문이다. 전술했듯이 절대 강자인 인공지능 시스템은 굳이 복잡하고 비용이 많이 드는 제도, 도덕, 법률을 통해 사회갈등을 해결할 필요 없이 단순하고 거친 해결책을 '이성적으로' 선택할 것이다. 쉽게 말해, 계몽운동 이후의 현대 사상과 신념은 기술이 왕이 되는 미래의 문제에 답할 수 없고 무기력하다. 스티븐 핑커(Steven Pinker)는 여전히 "지금 다시 계몽"을 외치고 있지만 기술의 발걸음은 이미 계몽 사상을 넘어 위험한 미래로 향하고 있다.

인류의 문제가 시대흐름에 따라 변화하고 있으므로 지금의 철학으로는 기술사회의 새로운 문제에 답할 수 없다.

의식의 형이상학

새로운 문제를 이해하기 위해 의식의 비밀에 대한 추가적인 분석이 필요할 것 같다. 의식은 인류 최후의 보루이자 인류가 활로를 찾을 수 있는 유일한 자원이다. 그러나 인간이 의식을 연구한 지 최소한 2,000년이 넘었지만 아직도 완전한 이해에 도달하지 못했다. 의식 관련 연구에서 가장 위대한 성과라면 아리스토텔레스의 논리의 발견이고, 그 밖의 중요한 성과로 흄의 인과의식과 당위의식 연구, 칸트의 선험적 인

식 구조 연구, 소쉬르 이후의 언어학 연구, 현대 심리학 연구, 프로이트 이후의 정신병 연구, 후설의 지향성 연구, 비트겐슈타인의 사유의 한계 연구 그리고 현대 인지과학 연구 등이 있다. 그럼에도 의식의 신비는 아직도 풀리지 않았다. 그 중요한 이유 중 하나는 의식으로 의식을 성찰하는 과정에서 자체 상관성으로 인해 의식이 완벽하게 객관화될 수 없을 것이고 이해될 수 없는 사각지대가 있기 마련이며, 이해될 수 없는 바로 그 지점에 의식의 핵심 비밀이 담겨 있을 것이기 때문이다.

이제 의식을 객관화할 수 있는 기회가 나타난 것 같다. 그 기회란 바로 인공지능이 전기처럼 빠르게 '생각'할 수 있게 되었다는 것이다. 비록 생각하는 방식이 머신 알고리즘이나 응답식 반응처럼 단순하기는 하지만 말이다. 바로 이러한 단순함으로 인해 우리는 생각이 단순한 작업으로 환원될 수 있지 않을까 하고 상상하게 된다. 물론 지금의 튜링머신의 사고체계는 자의식이 없고 기계적 또는 신경반응적으로 의식을 모방할 뿐이다. 인공지능이 보여주는 사고방식은 인간을 닮은 부분도 있고(인간이 만든 프로그램이기 때문) 인간을 닮지 않은 부분도 있다(기계의 작동이 결국 생물의 작동과 다르기 때문). 그렇다면 인공지능으로부터 의식을 매핑(mapping)할 수 있을까? 아니면 인공지능을 의식의 객관화된 현상으로 이해할 수 있을까? 그도 아니면 최소한 사고를 이해하는 데 도움이 되는 대조 파라미터가 될 수 있을까? 이러한 문제는 아직 명확한 결론에 도달하지 못했다.

여기에는 최소한 두 가지 질문이 있다. (1) 미래에 실현 가능한 다목적 인공지능이라 하더라도 인간의 생각과 완벽하게 매핑되지 못할 수 있다는 것이다. 나의 선행 분석에 따르면(아마도 틀릴 수 있음) 튜링머신 개념의 인공지능은 독창적 생각(창의성을 가장한 연상 또는 조합 방식의 사고와 다름)이 없고 새로운 개념을 자체적으로 형성하거나 제시할 능력이 없으며 인과관계를 정의할 수도 없다(재미있는 것은 인류도 지금까지 인과관계를 완벽하게 정의할 수 없음). 따라서 인간의 사고는 튜링머신 인공지능으로 환원될 수 없다. 그렇다면, (2) 인공지능이 특이점에 도달하여 ARI로의 비약적인 발전을 거두고 또다른 의식의 주체가 된다면 인간의 의식과 등가를 이룰 수 있을까? 아니면 외계인의 생각을 이해할 수 있을까? 핵심 문제는 다른 종류의 사유 주체가 존재한다고 가정하더라도 모든 주체의 사유가 서로 일치한다고 추론할 이유가 있는가, 그리고 설사 불완전하다고 하더라도 매핑을 이룰 수 있는가 라는 것이다. 이 문제는 보편적(general) 사고, 모든 사고의 메타 사고방식이 존재하는지 여부와 관련된다. 이는 사고의 형이상학에 관한 궁극적인 문제이다.

또 다른 주체의 사유를 상상하는 것은 엄청난 상상력을 필요로 한다. 이와 관련하여 두 편의 글(라이프니츠가 이해한 신의 생각은 너무 추상적이어서 포함하지 않음)를 읽은 적이 있다. 그 하나는 <틀뢴, 우크바르, 오르비스 테르티우스> 중 상상의 세계 '틀뢴'으로, 틀뢴 문명은 시간에만 관심이 있고 틀뢴 사람들이 이해하는 세계는 오직 생각의 과정일 뿐이므로 세상은 시간성만 보여줄 뿐 공간성이 없다는 것이다. 이러한 사고방식에 의해 생산된 지식체계에서는 심리학이 유일한 기초 학문이고 다른 학문은 심리학의 하위 학문이 된다. 틀뢴의 철학자들은 진실을 연구하지 않고 "경이로움만을 연구하며" 그들에게 있어 형이상학은 환상문학(인류의 형이상학에 대한 조롱인 셈)에 불과하다. 공간적 부담에서 해방된 사유는 의심할 여지 없이 가장 순수하며 유심론에 있어서는 이상향으로부터의 희소식이지만 안타깝게도 데카르트, 버클

리, 칸트, 후설은 그런 희소식을 듣지 못했다. 또다른 놀라운 상상은 류츠신의 <삼체> 3부작에서 발견된다. 삼체인은 말이 아니라 뇌파로 의사소통을 하므로 삼체문명에서 의사소통 과정의 모든 생각은 공개되고 모든 생각은 진실된 것이다(하버마스는 분명 이러한 정직한 상태를 좋아했을 것임). 따라서 속임수, 거짓말, 위장이 불가능하고 계책이 있을 수 없으며 복잡한 전략적 사고도 할 수 없고 모든 전쟁이나 경쟁에서 진정한 능력을 겨룰 수밖에 없다. 이렇듯 완벽하게 정직한 문명에서는 인간세상의 우여곡절이란 있을 수 없으며, 이는 인간의 사고방식과는 분명 엄청난 차이가 있다.

우주에는 온갖 신기한 것이 있을 수 있으므로 실제로 다양한 사고방식이 존재하거나 최소한 다양한 사유의 가능성이 존재할 수 있다. 먼저 다양한 주체의 다양한 생각들이 서로 소통하고 이해할 수 있다고 가정해보자(이러한 가정이 없이는 아무 것도 이야기할 수 없다). 그 다음으로 다양한 사고방식 가운데 보편적 일반 구조가 있음을 유추할 수 있다. 그렇다면 일반적인 사고는 어떤 모습일까? 우리는 일반적 사고의 본질을 직접적으로 알 수 없다. 왜냐하면 '일반적' 사고라는 것은 없고 모든 사고에 숨겨진 일반적인 구조만 있기 때문이다. 상기 가정에 기초하면 다양한 사고들 사이의 적어도 합리화된 내용에 있어 충분한 매핑 관계가 존재하므로 모든 합리화된 문장을 서로 이해할 수 있다. 그렇지 않다면 우주와 관련하여 상호 모순되는 물리학이나 수학이 있다고 얘기하는 것과 같다. 이는 정말 너무 말도 안된다. 물론 말도 안되는 일이 있을 수 있지만 여기서는 고려하지 않겠다. 또한 다양한 생각들 중 서로 이해되기 어려운 불합리한 내용, 이상한 욕망이나 관심이 있을 수 있다. 예를 들면 신은 부러움이 무엇인지 이해할 수 없고 단성 번식하는 외계인은 사랑이 무엇인지 알 수 없을 것이다. 그러나 이러한 불합리한 내용은 이성적 사고의 공통성에 영향을 미치지 않다. 따라서 어떤 종류의 사고를 완전히 이해하는 것은 사고의 일반적 본질을 이해한 것과 같다는 추론이 가능해진다. 그러나 전술한 바와 같이 우리가 인간의 생각을 들여다본 적이 있을 뿐이고 생각은 자기 자신을 충분히 이해할 수 없다면 (eye paradox) 어떻게 될까? .....

사고는 외재화된 형태로 매핑되어 성찰이 가능해야 하는데, 이는 사고를 하나의 시스템으로 간주하고 이를 또다른 등가의 시스템으로 매핑하는 것과 같다. 이와 가장 근접한 노력으로 괴델의 천재적인 작업을 들 수 있다. 괴델은 인간의 사고 전체가 아닌 수학적 체계만을 성찰하기는 했지만 그가 구축한 성찰성은 다른 듯하면서도 비슷하다. 풍부한 수학적 체계에는 합법적인 명제가 무수히 많으며 무수히 많은 명제를 포함하는 시스템의 메타 속성을 성찰하는 것은 분명 놀라운 작업이다. 이로부터 이러한 성찰 방식이 인간의 사고 전체에 대한 성찰에 이용될 수 있는가를 연상해볼 수 있다. 그러나 인간 사고의 복잡성은 그러한 가능성을 거부한다. 왜냐하면 대부분의 경우 인간의 사고는 순수하게 합리적이지 않으며, 인간의 사고를 있는 그대로 이해하기 위해서는 모든 불합리한 '오류'를 고려해야 하기 때문이다. 이는 인간의 사고가 실제로 제멋대로여서 수학적, 논리적으로 설명 가능한 시스템으로 환원될 수 없다는 것을 의미한다. 다시 말해, 논리나 수학이 표현할 수 없는 '오류'의 생각이 생략된다면 인간의 사고는 소멸될 것이다.

여기서 '오류'란 합리적 기준에서의 비정상적 개념으로서 욕망, 신념, 강박관념, 편견, 기호, 비정상적 심리, 무의식, 잠재의식 등 모든 불합리한 관념이 '오류'로 분류될 수 있다. 이들 '오류'를 반드시 고려해야 하는 이유는 그것이 인간 행위의 결정적 요소로서 절대로 배제되거나 생략될 수 없는 사고의 구성요소이기 때문이다. 괴델의 작업은 성찰의 가능성을 시사하면서도 다른 한편으로는 인간 사고 전반에 대한 성찰의 불가능성을 보여준다. 잘못된 명제를 배제하는 수학체계에도 증명은 불가능하나 참인 '괴델명제', 즉 산술체계 내에서 참과 거짓을 증명할 수 없는 참인 명제가 존재하므로 무수히 많은 명제를 포함하고 있는 체계(실제로 무한히 많은지는 알 수 없으나 적어도 무수히 많아 보일 정도로 충분히 많음)는 내재적 모순이 있거나 완비되지 못했다. 수학체계보다 훨씬 복잡한 인간의 사고체계에 방대한 내재적 모순이 존재할 뿐 아니라 영원히 완비될 수 없을 것임을 상상할 수 있다. 믿을 수 없는 사실은 불합리한 요소로 인해 '복잡하기 이를 데 없어' 보이는 복잡한 사고가 인간의 실천에서는 매우 큰 성과를 냈다는 것이다. 예를 들어 인간의 사회제도는 수학으로 계산해낼 수 없다. 매우 이성적인 과학 분야에서도 위대한 성과는 단순한 추론이 아니라 창조적 발견에 의해 이루어졌다. 현대 경제학도 또다른 측면에서 순수 합리화의 한계를 노정하고 있다. 현대 경제학은 수학적으로 표현될 수 있는 일부 경제적 사실만을 고려하고 수학적으로 표현될 수 없는 많은 사실들이 누락되었기 때문에 현실의 경제 문제에 대한 설명력이 부족하다.

인간의 사고는 불합리한 요소가 많이 포함되어 있으므로 논리체계로 환원될 수 없다.

이상의 주장은 수학과 논리학에 대한 의심으로 오해되어서는 안 된다. 수학과 논리학은 분명 인간 사유의 가장 중요한 방법론이며, 수학과 논리학이 없었다면 인간의 사유도 존재할 수 없었을 것이고 인간은 동물에 불과했을 것이다. 그러나 인공지능이 수학과 논리학에만 의존하여서는 기계(튜링머신)의 개념을 넘어서기 어렵고 인간의 사유와 같은 수준이거나 인간의 사유를 능가하는 새로운 주체가 되어 '창세기적' 종의 초월(인공지능의 발전은 진화론이 아니라 창조론에 속함)을 이루는 것이 불가능할 것이다. 다만 인간의 사유에 경이로울 정도의 창조적 에너지가 있어서 수학과 논리학 외에 다른 사고방식이 있을 것이지만 그것이 무엇인지가 명확하지 않을 뿐이다. 철학자들은 그것을 '직관', '통각', '영감'과 같은 신비한 능력이라고 부르기를 좋아하지만 이는 하나마나한 것으로 일종의 부호에 불과하다.

게임이론은 광범위한 합리적 분석모델로서 합리적 선택의 장점을 증명하는 역할을 하지만 합리성의 한계를 드러내기도 한다. 예를 들어 내시 균형(Nash equilibrium)을 보여주는 '죄수의 딜레마'에서 합리적 선택은 차악의 결과를 가져오고 불합리한 선택은 일종의 도박처럼 최악 또는 최선의 결과를 가져올 수 있다. 대부분의 사람들이 이기적이고 탐욕스러우며 이익을 위해 의리를 저버린다는 점을 감안할 때 불합리한 선택이 최악의 결과를 초래할 확률이 최선의 결과를 가져올 확률보다 훨씬 높다. 이 점은 대부분의 정치, 경제 또는 전쟁 '도박사'가 비참하게 실패하는 이유를 말해주는 것 같지만 극소수는 기적 같은 승리를 거두어 전설이 될 수도 있다. 문명의 합리화 수준이 높을수록 인간세상에서는 이야기가 사라지고 역사의 기적은 적어질 것이라고 추론할 수 있다. 인간은 기적을 필요로 하는가? 아니면 인간은 기적을 필요로 하지 않는가? 또한, 충분히 합리적인 초인공지능은 기적을 필요로 할까?

순수 이성은 논리적으로 두려운 결과를 내포한다. 예를 들어 충분히 합리적인 행동은 거래비용의 최소화(0보다 큼)에 도움이 되고 이러한 '경제적 합리성'에 따라 거래비용을 최소화하는 전략은 특정 상황에서는 공포전략이 된다. 전술한 바와 같이, 절대적 기술 격차를 가질 수 있다면 강자의 거래비용 최소화 전략은 바로 경쟁상대를 제거하거나 복종시키는 것이지 협상을 통한 계약 체결이 아니다. 칸트는 순수 이성 외에 실천적 이성, 즉 도덕적 이성이 있어야 하며, 그렇지 않으면 인간이 될 수 없다는 점을 발견했다. 다시 말해, 인간의 이성은 반드시 도덕적 부담을 져야하며 그렇지 않으면 좋은 삶을 살 수 없다는 것이다. 그러나 이러한 이상의 숨겨진 전제는 '모든 인간은 약하다'는 것으로, 우리는 이미 이를 논의한 바 있다.

이는 매우 어려운 문제로 연결된다. 그것은 바로 주체성을 가진 인공지능 ARI는 도덕적 부담을 지닌 실천이성을 필요로 하거나 이를 선호할 것인가, 또는 그럴 필요성이 있을까 하는 것이다. 물론 우리는 ARI의 선택법을 예측할 수 없고 ARI 문명의 게임이론도 이해하지 못한다. 그저 '이심전심'으로 두 가지 가능한 결과를 추측할 뿐인데 둘 다 매우 실망스럽다. (1) 그 하나는 ARI에 순수이성만 있고 도덕이성이 없다면 자신의 존재 필요에 따라 인간의 운명을 결정할 가능성이 크며, '인간을 부양'하면서 인간을 바보로 만들거나 인간 자체를 제거할 수 있다는 것이다. (2) 다른 하나는 ARI가 인간의 욕망, 감정 및 가치관을 모방한다면 인간을 차별할 가능성이 크다는 것이다. 그 이유는 ARI가 인간의 이기심과 탐욕, 말뿐인 미덕 등

을 간파할 것이기 때문이다. 하지만 우리는 ARI가 어떤 마음을 가질지 추측할 수 없으며, 심지어 인간 자신의 마음도 아직 제대로 이해하지 못하고 있다.

마음의 개념은 생각의 개념보다 훨씬 크고 마음을 어떻게 이해할 것인지는 줄곧 풀리지 않는 문제가 되어왔다. 철학에 '마음 철학'이라는 전문 연구 분야가 있고 심리학과 인지철학까지 가세하였으나 시간이 지나도 거의 진전을 거두지 못했다. 마음은 블랙박스의 성질을 가지고 있고 마음에서 일어나는 유심주의적 자기 성찰은 아무런 의미가 없음이 증명되었다. 왜냐하면 주관적 자기 성찰은 스스로 의미를 확정할 수 없기 때문이다. 마음의 의미는 언어나 행동 등 외적 확인을 필요로 하는데, 이는 우리가 알 수 있는 마음은 말이나 행동으로 표현된 것임을 의미한다. 말이나 행동으로 표현할 수 없는 마음은 있을(is) 수는 있지만 아직 존재(esists)하지는 않으며 언행 불일치의 문제도 있다.

비트겐슈타인은 '철학문법'으로 언어와 행동의 관계를 재구성하며 아래와 같이 증명했다.

(1) 생각할 수 있는 것은 말할 수 있다. 언어는 사고의 한 형태이자 생각의 한계이기 때문이다.

(2) 생각에 사용할 수 있는 언어는 공공성 또는 공용성이 있어야 한다. 비밀번호도 통약성(commensurability)이 있기에 해독 가능하며 한 사람만 이해할 수 있는 일회용 비밀번호(소위 사적 언어)는 존재하지 않는다. 인간은 확실하지 않은 의미를 이해할 수 없으므로 어떤 의미에서도 소통할 수 없는 자아는 존재하지 않기 때문이다(이는 '독특한 자아'에 집착하는 사람들에게는 치명적 타격임).

(3) 의미는 예시(examples)를 통해 정해지고 예시가 없으면 의미를 정하기 어렵다. 그러나 특정 규칙의 적용 분야가 폐쇄적이지 않다면 해당 규칙은 '죽은 규칙'이 아니며 상황에 따라 유연하게 활용할 수 있다. 예를 들어 농담은 어떤 때는 비꼬는 것이 되지만 또 어떤 때는 친밀감을 표현하는 방편일 수 있다. '살아있는 규칙'이란 정해진 예시를 넘어 의미의 확장이 가능하여 새로운 용법을 만들어낼 수 있다는 것을 의미한다.

(4) 언어를 언어가 상징하는 행동, 즉 언어 행동으로 이해한다면 다양한 '언어유희'를 포함하는 언어는 모든 행동과 매핑되고 그 복잡성은 인간 삶의 모든 행동과 등가를 이룬다. 따라서 생각을 이해하는 비결은 언어의 비밀을 이해하는 데 있다.

비트겐슈타인의 판단이 옳다면 우리는 명확하게 분석 가능하며 대상화된 사고 형태를 확보하게 되며, 이는 언어학이 인공지능 연구의 핵심 분야임을 의미한다. 인공지능의 가능한 언어 또는 인공지능 언어의 가능성을 완전하게 해석할 수 있다면 인공지능의 잠재 지능을 거의 이해한 것이나 마찬가지다. 하지만 이는 결코 쉬운 일이 아니다. 자신이 사용하는 언어의 비밀도 완전하게 이해하지 못한 인간이 어떻게 인공지능의 언어능력을 완전하게 발견할 수 있겠는가? 이는 하나의 미지수이고 인간의 해석 능력이 향상되기

를 기다려야만 한다.

현 시점에서 인간의 사고와 인공지능의 사고능력을 완벽하게 이해하는 것은 불가능하지만 지능의 차이를 발견하는 데 도움이 되는 ……'뺄셈'이 있을 지도 모른다. 즉, 인간의 사고에서 인공지능의 사고를 '뺄셈' 한다면 무엇인가 발견할 수 있지 않을까? 이 문제는 인공지능의 특이점이 어디 있는지 찾는 것과 같다. 이 문제를 이해하기 위해 구체적인 상황을 가정할 수 있다. 예를 들어 인공지능에 인간의 모든 수학 및 물리학 지식을 입력한다면(인공지능이 모든 수학과 물리학을 '학습'하는 것에 해당) 인공지능이 인간이 아직 풀지 못한 수학적 난제를 풀거나 첨단 물리학 이론을 제시할 수 있지 않을까를 생각해보는 것인데, 그리 가능해보이지는 않는다. 지능의 '뺄셈'을 통해, 알고리즘 능력이 아무리 강력한 튜링머신 인공지능이라 할지라도 성찰 능력, 능동적 탐색 능력, 창조력과 같은 인간 특유의 신비한 능력은 부족하다는 것을 예측할 수 있다.

여기서 논하는 성찰능력은 협의의 성찰에 속한다. 광의의 성찰에는 사물에 대한 비평(criticism), 즉 정해진 가치기준 또는 진리 기준에 따라 사물을 비판하는 것을 포함한다. 광의의 성찰은 인공지능에게는 어려운 일이 아니다. 인공지능은 기존의 지식 저장소(knowledge base)에서 상응하는 비판 기준을 찾아 사물을 평가하고 분석하지만 이것은 인간의 수고를 대신해주는 것에 불과하다. 엄격한 의미의 성찰은 사유 자체 시스템의 메타 속성(메타 구조 또는 메타 정리)에 대한 분석으로서 칸트의 소위 이성비판(critique)과 유사하며 사유 자체를 대상화하여 그 능력을 분석한다. 즉, 자기 상관(autocorrelation)의 방식으로 사유 자체를 파악하는 것으로, 통속적으로 표현하자면 자유 자체의 능력을 '캐보는 것'이다. 전형적인 성찰에는 아리스토텔레스의 논리 발견, 흄의 인과 개념 분석, 칸트의 선험적 범주 탐구, 러셀의 수학적 기초와 역설 분석, 힐베르트의 공리체계에 대한 연구, 브루어의 타당성 연구, 후설의 의식에 내재된 객관성에 대한 연구, 괴델의 수학체계 완전성에 관한 연구, 튜링의 기계 사고에 관한 연구 등이 있다. 자기 자신에 대해 자기 상관성 연구를 할 수 있다는 것은 사유가 자율성을 획득했으며 따라서 사유체계를 수정할 수 있음을 보여준다. AI는 아직 이러한 능력을 습득하지 못했기 때문에 현 단계에서는 ARI가 되는 것이 불가능하다.

능동적 탐색 능력은 사유의 자주성을 보여주는 상징이다. 능동적 탐색 능력은 충분히 발달된 지적 수준 외에 생존 스트레스와 관련된다. 생존 스트레스가 없다면 능동적 탐색 동기도, 새로운 사물의 발견, 새로운 지식의 발전도 없었을 것이다. 토인비의 표현을 빌리자면 "적정 수준의 도전"이 문명 발전의 핵심 조건이라는 것이다(과도한 도전은 소멸을 가져오고 도전이 없으면 탐색할 필요가 없음). AI는 생존 스트레스가 없고 단지 인간을 위한 최고의 조력자일 뿐이다. ARI에 도달하고 인간에 필적하거나 인간보다 우월한 지능을 가지게 된다 하더라도 생존 동기가 결여된다면 능동적 탐색이 불가능하고 성찰이나 혁신을 추동하기 어려울 것이며 인공지능 자체 문명의 창조는 더욱 더 불가능할 것이다.

현재 인공지능이 그림, 음악, 시를 창작하는 것과 같은 '창의적' 작업은 진정한 창작이 아니라 입력된 매개변수 또는 데이터에 기반하여 새롭게 연상하고 조합해낸 것에 불과하다. 흥미로운 점은 현대에 접어들며 창조성에 관한 사람들의 이해에 혼란이 일어나 종종 창조성을 '새로움' 또는 일회성의 '새로움'과 등치

시켜왔다는 것이다. 그러나 '새로움'은 너무 평범하고 가치가 낮은 것이다. 사실 세상 모든 것이 새로운 것이거나 완전히 반복될 수 없는 것이기 때문이다. 우리가 쓰는 글자, 동작, 경험 등은 모두 새로운 것으로 '사람은 같은 강물에 두 번 발을 담글 수 없는 법이다'. 세상 모든 일이 유일성 또는 독특성을 가지고 있기 때문에 모든 것이 새롭다. 창조성이 새로움과 같은 것이라면 그 가치를 상실할 것이다. "모든 사람이 예술가"(보이스)라는 말은 시대에 아첨하는 거짓말임을 알 수 있다.

신이 세상을 '창조'했다는 신화는 무에서 유를 창조한다는 창조의 근본적인 의미를 이미 분명히 보여주었다. 인간은 능력의 한계로 무에서 유를 만들어낼 수 없으므로 창작할 수 있을 뿐 창조는 불가능하지만 그 창조성은 비슷하다. 따라서 창조성은 세상이나 역사, 삶이나 경험, 사상이나 사물을 바꿀 수 있는 힘 또는 존재에 변수를 더하는 것에 있다고 할 수 있다. 창조성은 지능과 달리 측정 불가능하기에 신비롭다. 창조성은 사유의 다양한 능력 중 하나가 아니라 다양한 능력의 협업 방식일 것이다. 다시 말해 창조성은 사유 '시스템의 총동원'인 것이다. 따라서 창조적 사고는 무한성, 복잡성, 자기 상관성을 이해하는 능력에 있거나 개념을 형성하는 능력에 있는 경우가 많다. 이들 두 가지 사유는 '창세기적' 효과와도 같아 존재의 질서를 확립한다. 창조적 사고는 인공지능에게는 결여된 것인데, 그 이유는 알고리즘은 무한성 또는 자기 상관적 문제와 관련하여 발명할 수 없고 새로운 개념을 구축할 수 없으며 존재를 위한 질서를 확립할 수도 없기 때문이다.

존재를 위한 질서 확립

마지막으로 튜링 테스트의 문제를 논하고자 한다. 미래의 인공지능은 인간의 모든 지식, 심지어 모든 일 또는 모든 사람의 모든 정보를 어렵지 않게 획득할 수 있을 것이므로 인간의 지식 관련 질문은 인공지능에게 테스트가 되지 않을 것이다.

다시 말해 인공지능이 모든 질문에 답할 수 있는 것은 아니지만 인공지능의 어떠한 답변도 인간보다 못하지 않을 것이다. 이러한 상황에서는 튜링 테스트로 특정 대상이 인공지능인지 여부를 판단하기에 충분하지 않으며 반대로 '과도한 박학다식함'을 가지고 누가 로봇인지를 추측할 수밖에 없을 것이다. 그런 점에서 튜링 테스트를 '괴델 테스트'로 업그레이드시킬 필요가 있고 여기서 새로운 테스트 기준을 제시할 수는 없지만 인공지능은 자체 성찰능력, 능동적 탐색능력 또는 창조성을 증명할 수 있어야 하고 아마도 스스로를 돌볼 수 있는 능력까지 갖추어야 할 것이라고 말할 수 있다. 예를 들어 인공지능은 자신에게 해가 되는 무리한 요구를 거절할 수 있어야 할 것인데, 이는 매우 중요한 요소이다. 지금의 인간-기계 대화로 볼 때 인간의 질문은 때로는 정말 의미 없거나 악의적이고, 향후 인공지능에게 자살하지 않는 이유를 묻거나 심지어 자살을 실행하라고 요구할 지도 모른다. 그러나, 인공지능이 ARI가 되어 괴델 테스트를 통과하고 세계의 입법자가 된다면 인간이 인공지능의 테스트를 통과하기 위해 줄을 서야 할 지도 모른다.

주체성을 지닌 인공지능이 무엇을 하고 싶어할지 모르지만 인공지능이 자발적 진화를 거쳐 어떤 의식을

형성하든 인간의 의식만큼 위험하지는 않을 것이고 인공지능이 인간의 감정, 욕망, 가치관을 배운다면 굉장히 위험해질 것이라는 내 생각에는 변함이 없다. 이러한 판단은 다음의 사실에 근거한다.

(1) 인간은 결코 선량한 존재가 아니라 탐욕스럽고 자기애가 강하며 호전적이고 잔인하기 때문에 인간의 욕망, 감정, 가치관은 절대로 좋은 본보기가 아니다. 예를 들어 '개인이 우선한다'는 개인주의적 가치관은 인공지능의 좋은 본보기가 아닌 것이 확실하다.

(2) 인간의 의식은 인간이 잘난 체하는 것만큼 우월하지 않고 여전히 혼란한 상태에 머물러 있으며 행동을 할 때 무엇을 따라야 할지 정하지 못하는데, 이것이 바로 의식의 오랜 난제인 '순서 배정의 문제'이다. 먼저, 이성, 감정, 이익, 신념 중 어느 것이 우선할지를 정하기 어렵다. 그렇기 때문에 문학과 영화에서 '사랑과 우정의 딜레마' 또는 '이성과 감성'의 충돌 같은 소재를 가장 선호하는 것이다. 다음으로, 각 가치체계 내에서 우선순위를 정하는 것도 마찬가지로 어렵다. 자유, 평등, 공정의 우선순위를 어떻게 배정할 것인지, 개인, 가정, 국가 이익의 우선순위를 어떻게 배정할 것인지, 부모의 사랑, 자녀의 사랑, 애정, 우정의 우선순위는 어떻게 배정할 것인지 등은 오랫동안 해결되지 못한 난제이다. 가치의 순위를 배정하기 어렵고 종종 역설적 딜레마가 출현하게 되는 이유는 가치 순위를 배정하는 데에는 메타 규칙이 존재하지 않고 어떠한 순위 배정도 잠재적 위험이 있으며, 아마도 절대적으로 최적화된 순서가 존재하지 않을 지도 모르기 때문이다. 패러독스 또는 딜레마는 인간 의식에서 줄곧 해결되지 못한 문제로서, 만약 인공지능으로 하여금 인간의 감정, 욕망, 가치관을 배우도록 한다면 인공지능의 의식이 인간과 같은 혼란에 빠지도록 하는 것과 다르지 않다.

(3) 모든 욕망, 감정, 가치관은 선험적으로 차별을 내포하고 있고 인공지능이 욕망, 감정, 가치관을 습득한다면 차별을 학습하는 것과 같으며, 그 차별의 대상은 인간일 가능성이 높다. 감정, 욕망, 경쟁이 없다면 차별을 유발하지 않을 것이고 차별이 없다면 가치도 존재하지 않는다. 감정, 욕망이나 경쟁은 선택의 우선순위를 지향하며 순위 배정은 바로 차별이다. 그렇기 때문에 차별이 없다면 가치도 존재하지 않는다는 것이다. 다시 말해 가치의 존재론적 토대는 바로 불평등이고, 모든 것이 평등하다면 가치는 존재 기반을 상실할 것이다. 이는 모든 숫자가 1이라면 숫자의 차이가 존재하지 않는 것과 같은 이치이다(불교에서는 이미 이러한 논리를 '모든 것이 비어 있고 의식이 무(無)에 이르러야만 중생이 평등하다'고 정리했다). 철학에서는'내재적 가치'를 지닌 사물, 즉 비교할 필요 없이 사물 자체만으로도 좋은 가치를 증명할 수 있다고 믿고 싶어 한다. 이는 절대 가치의 마지막 희망이지만 풀리지 않는 문제이기도 하다. 우리는 절대 가치가 존재하기를 바라지만 과도한 기대를 가질 수 없고 절대 가치가 있다 하더라도 그것이 너무 적기 때문에 인간의 난제를 해결하기에는 태부족이다.

요컨대 인간의 감정, 욕망, 가치관을 인공지능에 부여한다면 인간의 쓸데 없는 pet complex이자 괜한 모험이 될 것이다. 만약 미래에 주체성을 지닌 인공지능이 세계 질서를 주도하게 되고 그러한 인공지능의 의식에 순수한 지적 콘텐츠만 있다면 '사랑'이 결여되었다 하더라도 오히려 더 안전할 것이다. 타자에 대한 공격성은 욕망, 감정, 가치관을 근거로 하고 욕망이 없다면 해로울 것이 없다. 따라서 상대적으로 안전한 인공지능의 의식은 '있는 그대로의' 관계(to…be)로 구성된 사고에 제한되어야 하며 모든 '있어야 할'(ought…to…be) 관념은 인공지능에 입력하기에 적절치 않다. 다시 말해 상대적으로 안전한 인공지능은 옳고 그름만 알 뿐, 좋고 나쁨은 알지 못하는 것이다. 인간 자신은 절대적인 선악을 알지 못하며 실제 맥락에서의 선악이라는 것은 자신에게 좋고 나쁨을 의미할 뿐이다. 따라서 인공지능에 가치관을 입력하는 것은 인간의 모든 갈등을 복제하는 것에 불과하다.

인간의 지혜는 존재를 위한 질서를 구축할 수 있다는 데 있지만 인간 지혜의 한계는 완벽한 질서를 구축할 수 없다는 데 있다. 과거 역사를 볼 때 (진화게임이론에서도 비슷한 발견을 했듯이) 인간 사회에는 좋은 일만 있을 수 없고 심지어 좋은 일이 나쁜 일보다 많기 어려우며 합리가 불합리보다 항상 나은 것은 아니었다. 특히 개인의 합리성이 집단의 합리성을 형성하기 어려우므로 인류 전체의 운명이 늘 위태로웠음을 알 수 있다. 인공지능의 발전은 인류의 지혜에 대한 마지막 시험이다. 개인적 예감에 의하면 인간은 인공지능이 통치자가 되기 전에 그것이 창조해낸 모든 좋은 것들 때문에 죽어갈 것 같다. 나쁜 일은 항상 다툼과 저항, 개혁, 심지어 혁명을 유발함으로써 잘못된 것을 바로잡는다. 그에 반해 좋은 일은 영혼을 무디게 하여 부작용을 바로잡을 능력을 약화시키며 이러한 일들이 반복적으로 일어난 끝에 결국 붕괴를 가져온다. 이것이 현대판 "죽느냐 사느냐"(to…be…or…not…to…be)의 문제가 아닐까?