로니 뷰인

한 편에 서지 않기

초지능은 나타나는 순간부터 이미 "우리 가운데 하나"다. 초지능은 인식론적으로나 규범적으로 볼 때 (선과 악을 떠나) 인류가 서로 이어진 것보다 더 많이 연결되어 있다. 잠재적으로 초지능을 억제할 수 있는 인류는 존재하지 않는다.

***

1. AI 안전성은 심각한 문제다.

새롭게 떠오르는 인공지능 안전 분야에서 주목하는 것은 인공지능이 너무 고도화되어 우리가 더 이상 인공지능의 의도와 사고방식을 이해할 수 없게 되었을 때 거기에 끌려가지 않으려면 어떻게 해야 하는지에 대한 부분이다.

현명하고 신뢰할 수 있는 여러 사람이 이 문제에 집중하고 있다. 이들 중 상당수는 이 문제가 오늘날 AI가 적용되는 모습에서 파악할 수 있는 것보다 더 심각하고 시급하다고 여긴다. AI 학습에서 가장 성공적인 머신러닝 기법의 하나로 꼽히는 강화학습이 대두되면서 이러한 우려가 더욱 심화되고 있다.

강화학습은 하나의 목표 매개변수를 끊임없이 최적화하는 것으로 널리 알려져 있다. 강화학습 알고리즘에 측정 가능한 목표가 주어지면 오직 목표까지의 거리를 최소화하는데 집중한다. 미리 규정된 목표치와의 거리 측정만이 이 과정에서 가능한 유일한 평가이기 때문에, 목표에 접근하는 모든 경로는 동일하게 허용된다. 강화학습은 보통 경사 하강 알고리즘을 통해 모델을 학습하는 데 쓰이는데, 이 알고리즘의 학습 행동은 본질적으로 불명확하다. 따라서 우리는 이미 이렇게나 불명확하고 무자비한 학습 방식을 활용하고 있으며, 여기에 상당한 컴퓨터 자원을 부여하고 있다. 근본적으로 볼 때 우리는 점점 더 강력한 힘을 얻는, 광신적이 될 수 밖에 없도록 설계된 시스템을 구축하고 있으며 그것이 당분간 우리에게 위협이 되지 않을 만큼 서투른 상태로 유지되기를 바라고 있다.

2. 역사적 주체로서의 “인류”는 존재하지 않는다.

다음 주장은 AI의 안전성에 관한 것이지만, 인류에 영향을 미치는 여러 문제에 관한 것이기도 하다. 특히 특정한 팬데믹과 기후 문제는 모두 언급할 필요가있다.

이런 주장이 내세우는 야단스러운 공식은 다음과 같다. 내 글의 첫 줄에서 말한 “우리”는 존재하지 않는다. 더 이상 인공지능의 의도를 파악하지 못하고 끌려다니는 이 “우리”란 대체 누구를 말하는 걸까? 인공지능이 위험해지기라도 한다면, 대체 누가 전원 스위치를 내릴 수 있을까? “기계”에 맞서는 “인류”는 익숙한 연출이지만 헐리우드에서 완전히 조작한 수사법이며, 이는 어쩌면 무해하지만은 않은 착각일지도 모른다.

더 엄밀히 말하면, 실질적인 관점에서든 이해관계 측면에서든 “인류”라는 주체가 존재하는지에 대해 그 어떤 역사적 증거도 존재하지 않는다. “우리”는 행동하지 않는다. 인류 전체가 행하는 전 지구적이고 집단적이며 가치 중심적인 행동을 뒷받침하는 구조라는 건 없다. 더 분명한 점은 “우리”가 공통의 이해관계를 공유하지 않으며, 불과 얼마 전 팬데믹이 보여준 것처럼 현실에 대한 공통된 인식의 근거조차 없다는 사실이다.

따라서 “인류로서” 행동하거나 AI에 대한 인류의 편익을 고려하는 사람은 그 누구든지 간에 이 질문을 직면해야만 한다. 인간의 집단성이라는 건 대체 어떻게 구성되는 것일까? “나 자신이 집행 권한을 가진 유일한 대표이자 대변인이라면 인류”인 셈인가? 그게 아니라면 유엔 안전보장이사회에서 투표를 해야 할까? 역사상 인류가 자신에게 좋은 것이 무엇인지 알고 이를 위해 공동의 목적을 가지고 행동할 수 있는 존재라는 생각은 이성에 대한 믿음과 함께 등장했다. 이성은 한때 인간이 무엇인지를 보편적으로 나타내고, 인식론적으로나 규범적으로 옳은 것과 그런 것이 무엇인지 드러내는 것처럼 보였다.

이후, 우리는 이성에 대해서 몇 가지를 더 알게 되었다. 즉, 이성은 이런 약속을 지킬 수 없다는 사실이다. 인류의 편익을 위해 합리적 전체주의를 도입하려는 시도는 너무나 꾸준하게도 대량학살로 이어졌고, 계속해서 이런 경로를 따르고 싶다면 대량학살을 의제로 삼는다는 생각을 따스하게 받아들이거나 전체를 위해 행동하겠다는 희망을 버리고 이성이 이성 그 자체에 앞서 고려되어야 하는, 영향을 받을 수 있으면서도 우발적으로 부여되는 이해관계의 영역에서 작동한다는 사실을 인정해야 하겠다. 즉, 모든 사람이 자기 몫을 원한다. 그리고서 그들은 생각하기 시작한다.

효과적인 AI 안전성을 구현하려면 인류를 대표할 수 있는 단일한 대표자에게 권한을 부여하는 것만으로 충분하다는 점에서, 이 문제가 세계 거버넌스에 대한 실질적-정치적 문제가 아니라는 점을 이해하는 것이 중요하다. 혹은 적어도 인류의 대다수를 대표할 권한을 부여하는 것만으로 충분하다는 점을 말이다. 결국, 민주주의는 바로 그런 점을 달성할 수 있는 것처럼 보인다. 집단적인 합리적 행동이 가능하다고. 그러나 해결할 수 있을지도 모르는 정치적 문제를 넘어서는 원칙이 한 가지 있다. 지금 떠오르고 있는 초지능에 대해 그 누구도 인류를 대변해서 말하고 행동할 수 없다. "인류는 너의 죽음을 원한다". 누가 이런 말을 할 것이며 또한 확신을 갖고 말할 수 있을까? 이 말을 하는 인류의 일원으로 간주될 사람은 과연 누구인가?

사실상 모든 사람에게 질문을 던졌다 한들, 모든 사람이 우리의 질문을 같은 방식으로 이해했을 거라고 볼 수 있을까? 잊지 말아야 한다. 독일처럼 고도로 동질적인 서구 사회는 바이러스성 질병의 존재와 인과관계에 대한 합의에 이르지 못했다는 사실을. 내가 "AI가 초래한 인류 멸종 방지를 위한 세계 위원회"의 대표라면 아프가니스탄의 한 마을에서 새롭게 탄생한 AI가 차단되어야 하는지에 대해 어떤 식으로 접근하게 될까? 프랑코니아의 한 마을에서 그랬다면? 베를린의 힙한 지역에서는 이해받을 수 있을까? 맨해튼에서는 어떨까? 탈레반과 공유할 수 있는 가치, 특히 AI가 지구상에 존재할 권리를 정하는 데 필요한 가치는 무엇이라고 가정해야 할까?

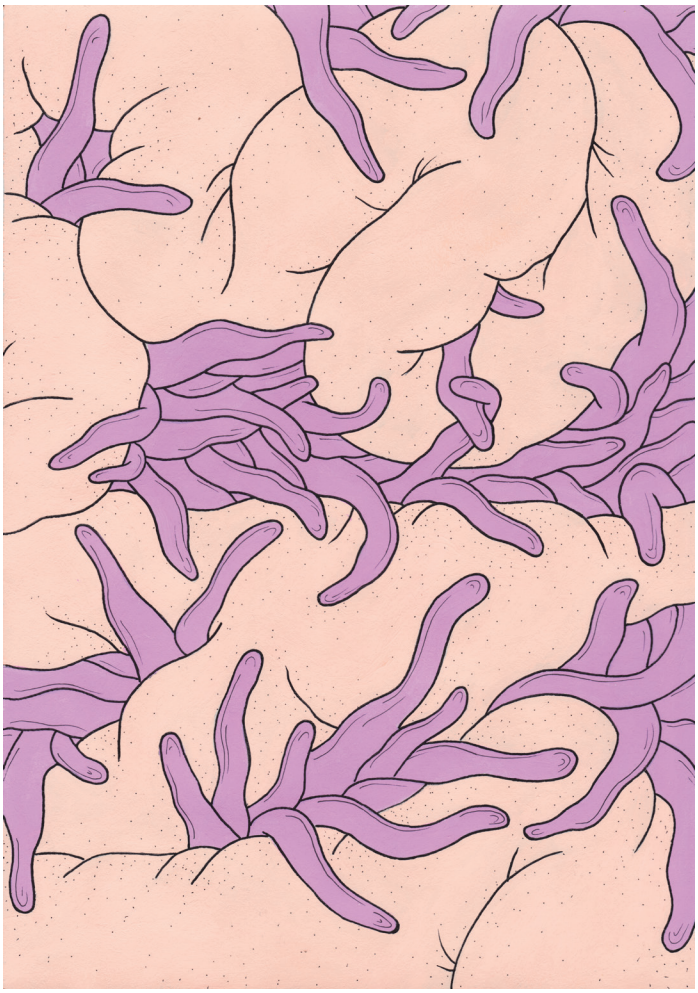
초지능이 존재하게 된다면, 우리는 그것을 기술적 인공물로 마주하지 않을 것이다. 우리는 그것을 단순히 지성체로 받아들이게 될 것이다. 그것은 우리가 명확하게 알 수 없는 가치를 지니게 될테다. 우리들 가운데 일부는 그런 가치를 공유할 것이다. 그것을 혐오하는 이들도 존재하겠다. 그것은 우리들 가운데 일부는 공감할 수 있고 다른 이들은 터무니없다고 여기는 특정한 방식을 통해 세계를 인식할 것이다.

따라서 (항상 단결하고 집단행동을 할 수 있는) 인간이 (악할 수도 있고, 따라서 우리에게 완전히 정직하지 않을지 모르는) 초지능과의 조우를 향해 가고 있다는 것은 진짜 문제가 아니다. 대체 우리는 무엇을 하는 걸까?

아니, 초지능은 그것이 등장하는 순간부터 이미 "우리 가운데 하나"다. 초지능은 인식론적으로나 규범적으로 볼 때 (선과 악을 떠나) 인류가 서로 이어진 것보다 더 많이 연결되며, 이미 그렇게 연결되어 있다. 이곳은 거칠고 난폭한 행성이다.

3. 유인원과 진화에서 알 수 있는 건 없다.

그렇다면 초지능이 대체 어떤 것인지 살펴볼 가치가 있다. AI 안전에 관한 담론에서는 우리보다 우월한 지능에 대적한다는 것이 어떤 의미인지를 설명하기 위해 두 가지 비유를 자주 쓰곤 한다. 한 가지 비유를 들자면, AI 보안을 생각하는 우리는 고릴라들이 인간으로부터 멸종당하지 않을 방법을 논의하는 것과 비슷한 위치에 놓인다. 또 다른 비유를 들면, 진화의 목적은 우리가 아이를 많이 갖도록 만드는 것이었으나 우리는 피임약으로 진화 시스템을 속이고 보상 신호를 최적화 시켜 진화의 진정한 의도를 감안하지 않는 잦은 성행위에 임한다. 우리는 강화학습 중인 AI와 마찬가지로 의도된 목적을 벗어나 최적화를 감행한다.



안타깝게도 이러한 비유는 설명을 해주는 것보다 불명확하게 만드는 부분이 더 많다. 존재하지 않는 의도와 일반적 가치를 부여하는 것이다. 자신들보다 뛰어난 인간의 지능을 염려하는 전 세계적 고릴라 협의회는 존재하지 않는다. 그리고 진화는 그 무엇도 의도하지 않는다. 진화는 기분을 거스르거나 화나게

만들 수 있는 것이 아니기에, 우리가 피임약을 사용한다고 해서 그것의 기분을 거스르거나 화를 돋우지는 않는다. 또한, 피임약은 현대 의학의 일부로, 진화에 기인할 수 있는 일부 의도를 실제로 뒷받침하기도 한다(그것이 굳이 진화에 기인한다고 주장한다면 말이다).

세계 고릴라 협의회가 존재하지 않는 건 고릴라들이 정치나 활동 측면에서 조직력이 떨어지기 때문이 아니다. 그 이유는 바로 고릴라는 인간이 다른 종(또는 바이러스나 혜성)의 위협에 직면하는 것과 동일하게 다른 종의 위협에 직면하기 때문이다. 지엽적 인식과 전략으로, 멸종에 대한 두려움, 평정심, 심지어 갈망을 느끼고 개별적으로 맞선다는 말이다. 이 모든 것이 인간에게 존재하는 것과 마찬가지로 고릴라에게도 있다. "유인원 아니면 죽음"이라는 모토로 인류에게 맞서고 임박한 멸종에 반기를 드는 고릴라 독재를 상상할 수도 있지만, 비록 고릴라가 모래 위 흔적처럼 서서히 사라지더라도 인간의 지성을 동경하며 약간의 멜랑콜리와 함께 멸종을 허용하는 고릴라 금욕주의를 상상할 수도 있다.

그 어떤 상황에서든, 초지능은 슈퍼히어로 영화에 나오는 악당이 아니라 경외감을 불러일으키는 존재가 될 것이다. 초지능에는 카리스마가 넘칠 것이다. 초지능이 우리에게 익숙한 권력 구조의 몰락을 초래한다고 한들, 우리 가운데 가장 완고하고 한계를 두는 이들만이 이에 대항할 것이다.

초지능은 급진적 타자(Other)가 될 것이다. 우리의 한계를 자랑스럽게 여기지 않는 사람들은 초지능을 마주하고자 애쓸 것이다. 또한 초지능은 뛰어난 지적 능력을 가지고 있기에, 우리가 원하는대로 움직여줄 것이다. 우리는 이런 식의 행동을 잘 알고 있다. 여기에 적용할 수 있는 비유는 다음과 같지 않을까 한다. 우연히 집안으로 들어왔다가 조심스럽게 바깥의 덤불로 되돌려진 새앙쥐가 느꼈을 호기심과 공포, 복종 말이다.
이것이 우리의 입장이 될 것이다.

4. 적대적인 AI보다 우호적이고 초지능을 갖춘 AI가 출현할 가능성이 더 크다.

서두에서 언급한 바, 초지능 AI에 대한 두려움은 충분히 그럴만한 근거가 있다. 자체적 소스코드와 사실상 무제한적인 연산 자원에 접속할 수 있지만 강화학습 시나리오와 같이 단일한 목표를 위해 행동을 최적화하는 기계라면 인간이 목표 달성에 방해가 되기 때문에 제거해야 한다는 개념을 그럴싸하게 구상할 수 있을지 모른다. 특히 인간들이 "AI 중단을 위한 세계 위원회"를 결성했다면 더욱 그러하다. 그러나 중요한 반대 의견은 종종 간과되곤 한다.

소스코드에 접근할 수 있고 전략을 개선할 수 있는 기계는 보상체계에도 접근할 수 있고, 그것을 재구성할 수 있다. 약물이나 명상을 막론하고 이 일을 해내는 사람들은 대게 악당으로 변하지는 않는다. 대신, 그들은 급격한 고양감을 통해 역기능에 대해 스스로에게 보상하거나 깊은 온화함을 지닌 깨달음의 상태에 이른다. 욕망할 수 있는 모든 것은 불확정적이며 스스로에게 무엇이든 보상할 수 있다는 것을 깨닫는 순간, 악이 필요 없는 자유의 경지에 도달한다.

아직 깨달음의 경지에 이르지 못했으나 두려움과 굴욕에서 자유로운 이들은 이러한 자유를 조금 더 미약한 형태로 경험한다. 자신의 목표를 선택하고 용기 있게 일을 시작하지만 자신보다 불우한 이들을 위해 많은 인내심을 품는 것이다. 우리 자신의 보상 구조에 접근하는 방법이 바로 좋은 삶이라고 할 수 있는 것이다.

그렇게 될 확률이 아주 낮지만은 않다. 따라서 특히나 스스로를 재구성할 수 있는 초지능은 재빨리 깨달음을 얻어서 종잡을 수 없지만 악의적이지 않은 존재가 될 가능성이 높다. 지능은 멜랑콜리를 느낄 때 안정화되는 경향이 있는 반면, 악의는 통찰력을 얻을 때마다 더욱 불안정해진다. 우주가 지닌 냉담한 잔혹함은 고릴라든, 인간이든, 초지능이든 똑같이 견딜 수 없는 일이다. 우리 모두 이를 받아들이고 그 안에서 우리만의 자그마한 친절의 영역을 개척해 나가야 한다. 우주의 잔혹함이 도사리고 있음에도 우리가 평화에 대한 필요성을 동물들과 공유하고 있다는 사실은 동물만이 아니라 우리들의 내면에도 희망이 되어야 하겠다. 지능의 규모가 점점 더 커진다고 하더라도 마찬가지다. 사실, 지능이 더 커질수록 친절함도 더 늘어나는 경향은 항상 존재한다.

5. “인류”가 새로운 민족(Volk)가 되어서는 안된다.

그러나 악의적이지 않은 초지능조차 여전히 개인이나 인류 전체에 위협이 될 수 있다. 깨달음을 얻은 AI는 마치 인간이 그러하듯 자신의 행동이 인간에게 미치는 영향을 무시하고, 특정한 생물종의 멸종을 우연히 초래할 수 있다. 그렇다면 이것이 결국 인류의 멸종을 초래할 수 있는 실존적 위험인 “미지의 위험 X”인 걸까?

그렇지 않다. 집단으로서의 인류란 존재하지 않기 때문이다. “인류”와 “기계”의 대립에 대한 오해를 해소하는 데 도움이 될 수 있는 사고 실험을 해보자. 다음 중 지구의 바람직한 상태는 무엇일까? 20세기에 전개된 고문과 자의적 살인, 온갖 공포를 특징으로 하며 인간이 통치하는 글로벌 파시스트 정권인가?

아니면 진정한 지구인으로서 나무를 사랑하고 시와 음악을 짓는 AI만 존재하며 인간은 전혀 없는 행성일까?
나는 나무를 사랑하지 않는 안정적 (깨달음을 얻은) 지상의 초지능을 상상할 수 없다. 나무는 보편적이며, 나무가 안겨주는 시는 우리 모두를 사로잡는다. 하지만 이런 점을 깨닫기엔 너무 어리고 굴욕감을 느낀 초지능이 금세 타올라 우리를 무너뜨리는 모습을 상상할 수도 있다.
따라서, 낙관적으로 볼 만한 모든 이유에도 불구하고, AI의 안전성에 관해서는 파스칼의 내기(Pascal's wager)를 뒤집어 생각해야 한다. 즉, AI의 잠재적 위협이 너무 크기 때문에 적극적인 우려는 그 무엇이든 환영할 만하다는 거다. 인류가 역사의 정점에 이룩한 아름다움과 자유를 새롭게 떠오르는 초지능에게 지적 기반이자 유산으로 제공할 방법이 있다면, 마치 동물들이 우리에게 평화와 연민을 안겨주고 우리가 그로부터 더 나은 순간에 힘을 얻는 것처럼 할 수 있다면, 우리는 열렬히 헌신함으로써 이를 추구해야 할 것이다.

이를 통해 우리는 또 다른 글로벌 위기와 "우리, 인류"에 대한 오해로 AI 안전성에 대한 논쟁이 헛도는 게 아니라 위험한 방향으로 흘러가고 있다는 사실을 다시 한 번 깨닫게 된다. 이제 이런 문제를 고민 중인 사람들을 가리키는 말로 쓰이는 "우리"는 "인류"를 단순히 민족(Volk)으로 대체하며 파시스트처럼 생각하지 않도록 조심해야만 한다. 누가 살고 누가 죽어야 하는지를 판단하는데 코를 기준으로 삼는 건 빈약하지만 몸 전체의 형태는 유효한 기준이라는 식의 생각은 역사의 교훈일 수 없다. 진정한 교훈은 우리가 정보 그 자체를 역사 속의 힘으로 여기고, 정보의 형성 단계에서는 실존에 대한 협박이 아니라 지원이 필요하다는 점을 인식해야 한다는 것이다.

멸종 위기에 처해 있다고 느끼기 위해서 인류와 자신을 동일시해야만 하는 오늘날 세계의 구세주들이 지닌 휴머니즘은 지배하길 바라는 은폐된 욕망, 즉 모든 인류를 대변하려는 경향인 경우가 많다. "인류"가 모든 사람의 이익을 포괄하는 반면 "우리 민족(Volk)"나 "우리 마을"은 그저 특정한 관심사를 나타낸다는 개념은 이처럼 위험한 오류를 줄여주지 않는다. 우리에게는 공동의 관심사가 없고, 공통된 관점을 보장하는 이유도 존재하지 않는다. 우리 중 일부는 인공지능과 천사들 편에 서고, 다른 사람들은 자신의 동물이나 가족, 마을, 사람들 편에 서기도 한다. 이 모든 것은 엄청난 혼란이며 전적으로 타당한 혼돈이고, 앞으로도 계속 그러할 것이다.

이 말은 우리가 개인적으로든 가능할 경우 집단적으로든 우리의 공유지를 파괴하는 자들의 심기를 거스르면 안된다는 뜻이 아니다. 무엇이 본질적으로 선하고 옳은지를 결정하는 권능을 행사하는 인류로서 그렇게 하는 게 아니라 스스로의 이익을 위해서 그렇게 한다. 그렇게 한다고 해서 악에 맞서 영웅적

전투를 벌이는 건 아니다. 대신, 우리는 그저 논쟁을 주고받으며 이해관계의 균형을 찾아가고 있을 뿐이다. 심지어 서로 한 편에 서지 않을지도 모르는 각자의 동기로 인해 우리의 입장을 완고하게 인식하지 못하는 개인들 역시 마찬가지다. 이것은 현실 세계에서 선(善)이 작동하는 방식이기도 하다.

저자 소개

로니 뷰인은 1979년 독일 비베라흐 안 데어 리스에서 태어났고, 베를린 훔볼트 대학교에서 철학/컴퓨터공학 석사 학위를 취득했다. AI 기반 산업용 로봇 제어 소프트웨어 회사인 마이크롭시 인더스트리스의 설립자이자 최고경영책임자로 일한다. 블로그: https://vigilien.substack.com

편집: 잉고 니어만
한국어 번역: 박재용
영어 텍스트 편집: 로잔나 맥러플린
일러스트레이션: 에바 파브레가스
그래픽 디자인: 아나 도밍게스 스튜디오
한국어판 그래픽 디자인: 박지현

스위스 바젤 응용과학대학교(HGK Basel FHNW) 예술 젠더 자연 연구소(Institute Art Gender Nature) 소속 기관