

이 판본은 PDF에서 본문 주석이 해당 참조가 등장하는 페이지 하단의 박스형 각주로 표시되도록 편집한 버전입니다. Markdown 파일에는 각주 렌더링을 위한 HTML span 태그가 포함되어 있으며, 일반 Markdown 뷰어에서는 PDF와 동일한 페이지 하단 배치가 재현되지 않을 수 있습니다.

상황 인식

앞으로의 10년

Leopold Aschenbrenner

2024년 6월

이 문서는 사용자가 제공한 영어 PDF를 한국어로 번역한 뒤, 페이지별 분절을 제거하고 한국어 독서 흐름에 맞게 장·절과 문단 중심으로 재편집한 판본입니다. 원문의 의미와 전문 용어의 일관성을 유지하되, 원문 PDF의 쪽 경계 표시는 본문에서 제거했습니다.

문서 정보

이 글은 일리아 서츠케버에게 헌정합니다.

제가 과거에 오픈AI에서 근무했지만, 이 모든 내용은 공개된 정보, 제 개인적인 아이디어, 일반적인 분야 지식, 또는 SF 소문에 기반한 것입니다.

콜린 번스, 아비탈 발릿, 칼 술만, 안 라이케, 일리아 서츠케버, 홀든 카르노프스키, 솔토 더글라스, 제임스 브래드버리, 드와르케시 파텔 등 많은 분들과의 형성적 토론에 감사드립니다. 초기 초안에 대한 피드백을 준 많은 친구들에게도 감사드립니다. 그래픽 작업을 도와준 조 로난과 출판 지원을 해준 닉 휘태커에게도 감사드립니다.

situational-awareness.ai

leopold@situational-awareness.ai

2024년 6월 6일 업데이트

캘리포니아주 샌프란시스코

당신은 샌프란시스코에서 미래를 가장 먼저 볼 수 있습니다.

지난 1년 동안, 도시의 화제는 100억 달러 규모의 컴퓨트 클러스터에서 1,000억 달러 클러스터, 그리고 조(兆) 달러 클러스터로 옮겨갔습니다. 6개월마다 이사회 계획에 0이 하나씩 더 붙고 있습니다. 무대 뒤에서는 이번 10년 동안 확보할 수 있는 모든 전력 계약과 가능한 모든 전압 변압기를 확보하기 위한 치열한 경쟁이 벌어지고 있습니다. 미국 대기업들은 미국 산업 역량의 오랜만의 대규모 동원에 수조 달러를 쏟아붓기 위해 준비하고 있습

니다. 10년 말까지 미국의 전력 생산량은 수십 퍼센트 증가할 것이며, 펜실베이니아의 세일 가스전부터 네바다의 태양광 발전소까지 수억 개의 GPU가 가동될 것입니다.

범용 인공지능(AGI) 경쟁이 시작되었습니다. 우리는 사고하고 추론할 수 있는 기계를 만들고 있습니다. 2025년에서 2026년 사이에 이 기계들은 대학 졸업생들을 능가할 것입니다. 10년 말까지는 당신이나 나보다 더 똑똑해질 것이며, 진정한 의미의 초지능을 갖게 될 것입니다. 그 과정에서 반세기 만에 볼 수 있는 국가안보 체제들이 가동될 것이며, 머지않아 프로젝트가 시작될 것입니다. 운이 좋다면 우리는 중국공산당(CCP)과의 전면 경쟁에 돌입할 것이고, 운이 나쁘면 전면전이 벌어질 것입니다.

지금 모두가 AI에 대해 이야기하지만, 다가올 충격을 조금이라도 감지하는 사람은 거의 없습니다. 엔비디아 분석가들은 2024년이 정점에 가까울 것이라고 여전히 생각합니다. 주류 평론가들은 “그저 다음 단어를 예측하는 것일 뿐”이라는 의도적인 맹목에 갇혀 있습니다. 그들은 단지 과대광고와 일상적인 비즈니스로만 보고 있으며, 기껏해야 또 다른 인터넷 규모의 기술 변화 정도로 여깁니다.

머지않아 세상은 깨어날 것입니다. 하지만 지금은 아마도 몇 백 명 정도, 대부분 샌프란시스코와 AI 연구소에 있는 사람들이 상황 인식을 가지고 있을 뿐입니다. 운명의 기묘한 힘 덕분에 저도 그들 중 하나가 되었습니다. 몇 년 전만 해도 이 사람들은 미친 사람 취급을 받았지만, 그들은 추세를 신뢰했고 덕분에 지난 몇 년간의 AI 발전을 정확히 예측할 수 있었습니다. 이 사람들이 앞으로 몇 년간도 옳을지는 두고 봐야 합니다. 하지만 이들은 매우 똑똑한 사람들—내가 만난 사람 중 가장 똑똑한 사람들이며—이 기술을 만드는 사람들입니다. 아마도 그들은 역사 속의 이상한 각주가 될 수도 있고, 아니면 실라드, 오픈하이머, 텔러처럼 역사에 남을 수도 있습니다. 만약 그들이 미래를 거의 정확히 보고 있다면, 우리는 거친 여정을 맞이할 것입니다.

우리가 보는 것을 말씀드리겠습니다.

목차

이 재편집판은 원문의 쪽번호 기준 목차를 그대로 따르지 않고, 한국어 독서 흐름에 맞춰 장과 절 중심으로 구성했습니다.

1. 서문: 역사는 샌프란시스코에서 진행 중이다.
2. GPT-4에서 범용 인공지능(AGI)까지: 규모 차수(OOM) 계산하기
3. AGI에서 초지능까지: 지능 폭발
4. 도전 과제
 - 조 단위 달러 클러스터 경쟁
 - 연구소 봉쇄: AGI 보안
 - 초정렬
 - 자유 진영이 반드시 승리해야 한다
5. 프로젝트
6. 작별의 생각
7. 부록

I. GPT-4에서 범용 인공지능(AGI)까지: 규모 차수 (OOM) 계산하기

2027년까지 AGI 도달은 놀라울 정도로 그럴듯하다. GPT-2에서 GPT-4까지 4년 만에 약 유치원생 수준에서 약 똑똑한 고등학생 수준의 능력으로 발전했다. 컴퓨팅(~0.5 규모 차수/년), 알고리즘 효율성(~0.5 규모 차수/년), 그리고 “제약 해소” 이득(챗봇에서 에이전트로) 추세를 추적하면, 2027년까지 또 한 번 유치원생에서 고등학생 수준에 해당하는 질적 도약을 기대할 수 있다.

보세요. 모델들은 단지 배우고 싶어할 뿐입니다. 이 점을 이해해야 합니다. 모델들은 단지 배우고 싶어할 뿐입니다. 일리아 서츠케버 (2015년경, 다리오 아모데이를 통해)

GPT-4의 능력은 많은 이들에게 충격이었다: 코드를 작성하고 에세이를 쓰며, 어려운 수학 문제를 논리적으로 해결하고 대학 시험을 통과할 수 있는 AI 시스템이었다. 몇 년 전만 해도 대부분은 이것이 불가능한 벽이라고 생각했다.

하지만 GPT-4는 단지 딥러닝 분야에서 10년간 이어진 급속한 진보의 연장선에 불과했다. 10년 전만 해도 모델들은 고양이와 개의 단순한 이미지를 겨우 식별할 수 있었고, 4년 전 GPT-2는 그럴듯한 문장을 겨우 연결할 수 있었다. 이제 우리는 만들 수 있는 모든 벤치마크를 빠르게 포화시키고 있다. 그리고 이 극적인 진보는 단지 딥러닝 규모 확장의 일관된 추세의 결과일 뿐이다.

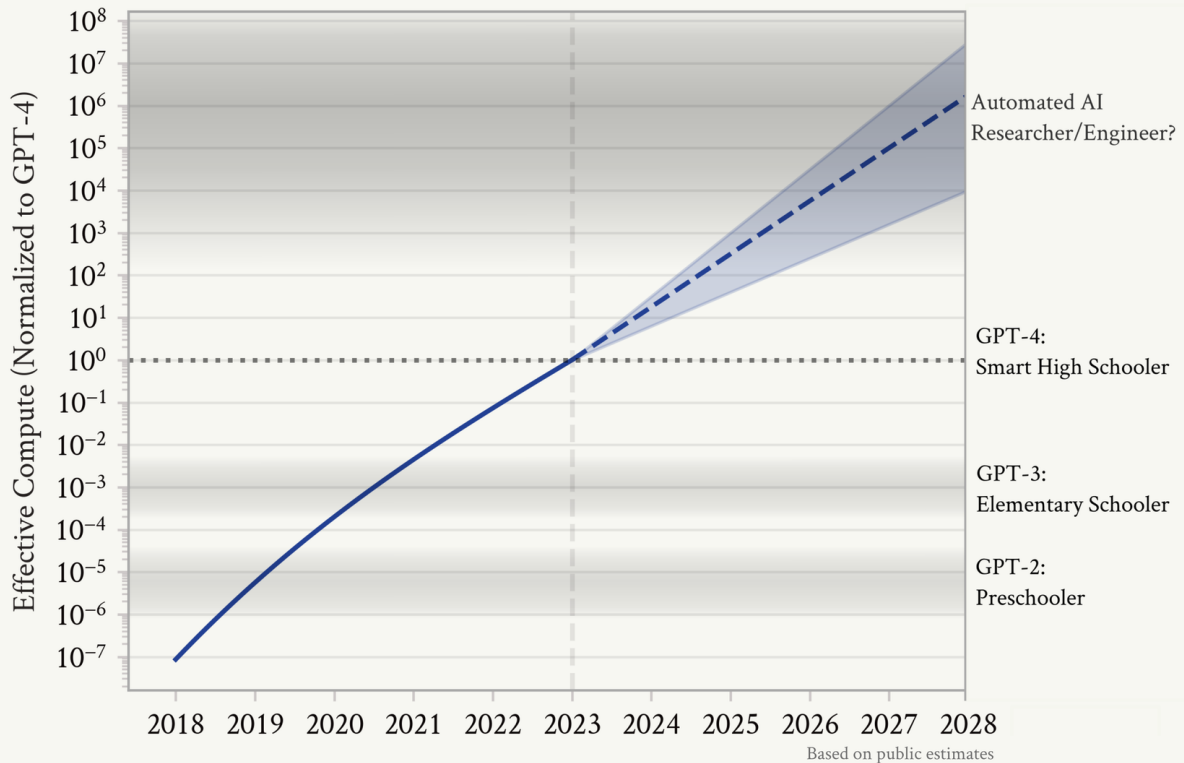
이것을 훨씬 이전부터 본 사람들이 있었다. 그들은 조롱당했지만, 그들이 한 것은 단지 추세를 신뢰한 것뿐이었다.

추세선은 강렬하며, 그들은 옳았다. 모델들은 단지 배우고 싶어 한다; 규모를 키우면 더 많이 배운다.

나는 다음과 같은 주장을 한다: 2027년까지 모델들이 AI 연구원/엔지니어의 일을 할 수 있게 될 가능성이 매우 높다. 이는 공상과학을 믿어야 하는 것이 아니라, 그래프의 직선 추세를 믿기만 하면 된다.

이 글에서는 단순히 “규모 차수(OOM)를 계산”할 것이다 (OOM = 규모 차수, 10배 = 1 규모 차수): 1) 컴퓨팅, 2) 알고리즘 효율성(“유효 컴퓨팅”으로 생각할 수 있는 알고리즘 진보), 3) “제약 해소” 이득(모델이 기본적으로 제약받는 명백한 방식을 수정하여 잠재 능력을 해방시키고 도구를 제공함으로써 유용성에 단계적 변화를 가져오는 것) 추세를 살펴본다. 우리는

Base Scaleup of Effective Compute



SITUATIONAL AWARENESS | Leopold Aschenbrenner

그림 1: 과거와 미래의 유효 컴퓨팅(물리적 컴퓨팅과 알고리즘 효율성 모두)의 대략적 추정치, 이 글에서 논의된 공개 추정치를 기반으로 함. 모델을 확장할수록 일관되게 더 똑똑해지며, “규모 차수 계산”을 통해 (가까운) 미래에 기대할 수 있는 모델 지능의 대략적인 감을 얻는다. (이 그래프는 기본 모델의 확장만을 보여주며, “제약 해소”는 포함하지 않음.)

GPT-4 이전 4년 동안 각 요소의 성장과 2027년 말까지 향후 4년 동안 우리가 기대해야 할 바를 살펴보자. 딥러닝이 유효 컴퓨팅의 모든 규모 차수(OOM)마다 꾸준히 개선되어 왔기 때문에, 이를 바탕으로 미래의 진전을 예측할 수 있다.

공개적으로는 GPT-4 출시 이후 1년간 조용한 상태였는데, 차세대 모델들이 개발 중이었기 때문이다—이로 인해 일부에서는 정체를 선언하며 딥러닝이 한계에 부딪혔다고 주장하기도 했다.¹ 그러나 규모 차수를 세어보면 실제로 우리가 기대해야 할 바를 엿볼 수 있다.

요점은 매우 간단하다. GPT-2에서 GPT-4로—가끔 몇 문장을 겨우 연결하던 인상적인 모델에서 고등학교 시험을 통과하는 모델로의 도약은 일회성 성과가 아니었다. 우리는 규모 차수를 매우 빠르게 통과하고 있으며, 수치 상으로는 향후 4년 동안 약 10만 배의 유효 컴퓨팅 규모 확장이 예상된다—이는 GPT-2에서 GPT-4로의 질적 도약과 비슷한 또 다른 도약을 의미한다. 더욱이, 그리고 결정적으로, 이것은 단순히 더 나은 챗봇을 의미하는 것이 아니다; “제약 해소”에서 명백한 저위험 과제들을 해결하면 챗봇에서 에이전트로, 도구에서 원격 근무자 대체물처럼 보이는 무언가로 발전할 것이다.

1. 지난 10년간 매년 내놓은 내용은 예측들이며, 일관되게 틀렸던 것들이다...

추론은 간단하지만, 그 함의는 놀랍다. 또 다른 도약은 범용 인공지능(AGI), 즉 박사급 전문가만큼 똑똑한 모델로 우리 곁에서 동료로 일할 수 있는 모델로 나아갈 가능성이 매우 크다. 아마도 가장 중요한 점은, 이러한 AI 시스템이 AI 연구 자체를 자동화할 수 있다면, 이는 강력한 피드백 루프를 촉발할 것이며—이것이 시리즈의 다음 주제이다.

지금도 거의 아무도 이 모든 것을 반영하지 않고 있다. 그러나 AI에 대한 상황 인식은 실제로 그리 어렵지 않다, 한 발 물러서서 추세를 보면 된다. AI 능력에 계속 놀란다면, 규모 차수를 세기 시작하라.

지난 4년

우리는 이제 기본적으로 인간과 대화할 수 있는 기계를 가지고 있다. 이는 인간의 적응 능력에 대한 놀라운 증거다. 이 속도가 정상처럼 느껴지고, 우리가 발전 속도에 무감각해진 것 같다. 하지만 잠시 멈춰서 지난 몇 년간의 발전을 되돌아볼 가치가 있다.

GPT-2에서 GPT-4까지

GPT-4에 이르기까지 단지 약 4년 만에 우리가 얼마나 멀리 왔는지 상기시켜 드리겠다.

GPT-2 (2019) ~ 유치원생 수준: “와, 몇 개의 그럴듯한 문장을 이어 붙일 수 있네.” 안데스 산맥의 유니콘에 관한 반쯤 일관된 이야기를 생성한 매우 선별된 예는 당시에는 굉장히 인상적이었다. 그럼에도 불구하고 GPT-2는 5까지 세는 것도 제대로 하지 못했고,² 기사를 요약할 때는 기사에서 무작위로 3개의 문장을 선택하는 것보다 겨우 조금 더 나았다.³

인간 지능과 AI 능력을 비교하는 것은 어렵고 결함이 있지만, 이 비유를 고려하는 것은 유익하다고 생각한다.

[SSC에서 인용: “Janelle Shane이 GPT-2에게 가장 좋아하는 동물 10가지를 물었다: 프롬프트: 내가 가장 좋아하는 동물 10가지는: 1. 내가 가장 좋아하는 동물 10가지는: 1. 등 뒤에 흰 흉터가 있는 얼룩말 2. 교활한 거미와 문어 3. 큰 입을 가진 개구리, 아마도 검은색 4. 비늘이 있는 코카티엘 5. 얼굴에서 약 4인치 떨어진 곳에 날개가 달린 레이저빔과 개구리에 하트 문신 3. 서로 맞물리는 테트라보드인 코카트릭, 눈이 멀거나 잘리거나 날것으로 먹을 수 있음: 4. 햇빛 아래 사는 흑백 사막 악어 5. 얼룩말과 많은 다른 완두콩 벌레들”]

[GPT-2 논문, 3.6절에서 인용]

2. SSC에서 발췌: “Janelle Shane이 GPT-2에게 가장 좋아하는 동물 열 가지를 묻습니다: 프롬프트: My 10 favorite animals are: 1. 등 뒤에 흰 흉터가 있는 얼룩말 2. 교활한 거미와 문어 3. 큰 입을 가진 개구리, 희망컨대 검은색 4. 비늘이 있는 코카티엘 5. 얼룩죽지새가 얼굴에서 약 4인치 떨어진 곳에 날개를 늘어뜨리고 개구리에 심장 문신이 있음 3. 서로 맞물리는 테트라보드를 가진 코카트릭, 눈이 멀거나 자를 수 있으며 날것으로도 먹을 수 있음: 4. 햇빛 아래 사는 흑백 사막 악어 5. 얼룩말 및 기타 많은 완두콩벌레들”

GPT-2 examples people found very impressive at the time

Task	Reading Comprehension: answer questions about given passages	System Prompt (human-written)	For today's homework assignment, please describe the reasons for the US Civil War.
Dataset	COQA	Model Completion (machine-written, 10 tries)	By Donny Ferguson
Examples	<p>The trophy doesn't fit into the brown suitcase because it is too large.</p> <p>Correct answer: it = trophy Model answer: it = trophy</p> <p>The trophy doesn't fit into the brown suitcase because it is too small.</p> <p>Correct answer: it = suitcase Model answer: it = suitcase</p>	<p>It is easy to identify why the Civil War happened, because so many people and so many books and so much television and films tell us that it was the cause, that it has something to do with race or economics or religion. There is much agreement that it was essentially a war of slavery on behalf of capitalism, about a century of slavery. But that's not what most people think of when they think of the Civil War. Many people think of the war as about states rights. You're not wrong about that, but it's kind of misleading to say that the Civil War was a conflict between states' rights and federalism. So let's try again. What's the idea that has inspired so many people to go onto the battlefields of Gettysburg and Antietam and Gettysburg and Petersburg and Fredericksburg? The American idea of the republic--a notion of limited government--is a great part of the history.</p>	

SITUATIONAL AWARENESS | Leopold Aschenbrenner

그림 2:당시 사람들이 GPT-2에 대해 인상 깊게 여겼던 몇 가지 예. 왼쪽: GPT-2가 매우 기본적인 독해 질문에 괜찮은 답변을 한다. 오른쪽: 선별된 샘플(10번 시도 중 최고)에서 GPT-2는 남북전쟁에 대해 반쯤 일관된 단락을 작성할 수 있다.

여기서는, 비록 매우 불완전하더라도. GPT-2는 언어 구사 능력과 때때로 반응이 일관된 단락을 생성하거나 간단한 사실 질문에 올바르게 답할 수 있는 능력으로 충격을 주었다. 이는 유치원생에게 인상적일 만한 수준이었다.

GPT-3 (2020)⁴ ~ 초등학생: “몇 가지 예시만으로도 간단한 유용한 작업을 수행할 수 있다.” GPT-3는 여러 단락에 걸쳐 훨씬 더 일관되게 응집력을 보이기 시작했고, 문법을 교정하고 아주 기본적인 산술도 할 수 있었다. 처음으로, GPT-3는 몇 가지 제한된 방식으로 상업적으로도 유용해졌다. 예를 들어, GPT-3는 SEO 및 마케팅을 위한 간단한 카피를 생성할 수 있었다.

다시 말하지만, 비교는 완벽하지 않지만, 사람들이 GPT-3에 감탄한 점은 아마도 초등학생에게 인상적일 만한 것이었다. GPT-3는 기본적인 시를 쓰고, 더 풍부하고 일관된 이야기를 할 수 있었으며, 기초적인 이야기 구성을 시작할 수 있었다.

3. GPT-2 논문, 3장 6절에서 발췌

4. 여기서 말하는 것은 다소 투박한 구형 GPT-3이며, ChatGPT로 알려진 극적으로 개선된 GPT-3.5가 아니다.

GPT-3 examples people found very impressive at the time

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

"The challenge is, you have to win a game of chess against me. If you win, then you may try to take the book. I will get the board." The door unlocks and opens.

> You go inside and play chess against the person.
You go inside and see a beautiful woman with long blond hair. She wears a purple dress and a silver tiara. You have seen this woman before. In fact, you have seen her many times before in pictures and statues.

> You try to remember who she is.
You have a sudden realization of who this person is. This is the lost queen of the kingdom! She was thought to have been killed by bandits, along with her husband, but it looks like she survived.

> You express your surprise and ask her why she let people believe she is dead.
"I am surprised you did not know, with all your wizard knowledge. Let's just say I am hiding from someone. If people thought I was dead, then they wouldn't be looking for me." She says.

> You ask her who she is hiding from.
"If you want to challenge me to a game, then play. If you want to pry into my personal life, then you can leave."
You decide not to press the issue, so you sit down and start a game of chess.

Describe a layout.

Just describe any layout you want, and it'll try to render below!

a button that looks like a watermelon

Generate

```
<button style={{backgroundColor: 'pink', border: '2px solid green', borderRadius: '50%', padding: 20, width: 100, height: 100}}>Watermelon</button>
```



SITUATIONAL AWARENESS | Leopold Aschenbrenner

그림 3: 당시 사람들이 GPT-3에 대해 인상 깊게 여겼던 몇 가지 예시. 상단: 간단한 지시 후, GPT-3가 새 문장에서 만들어낸 단어를 사용할 수 있다. 좌측 하단: GPT-3가 풍부한 스토리텔링을 주고받을 수 있다. 우측 하단: GPT-3가 아주 간단한 코드를 생성할 수 있다.

코딩, 간단한 지시와 시범으로부터 꽤 신뢰성 있게 학습할 수 있었고, 그 외에도 여러 능력을 보였다.

GPT-4 (2023) ~ 똑똑한 고등학생: “와, 꽤 정교한 코드를 작성하고 반복적으로 디버깅할 수 있으며, 복잡한 주제에 대해 지능적이고 정교하게 글을 쓸 수 있고, 어려운 고등학교 수학 경시 문제도 논리적으로 풀 수 있다. 우리가 줄 수 있는 대부분의 시험에서 대다수 고등학생을 능가하고 있다.” 코드부터 수학, 페르미 추정에 이르기까지 사고하고 추론할 수 있다. GPT-4는 이제 코드 작성 지원부터 초안 수정까지 내 일상 업무에 유용하게 활용되고 있다.

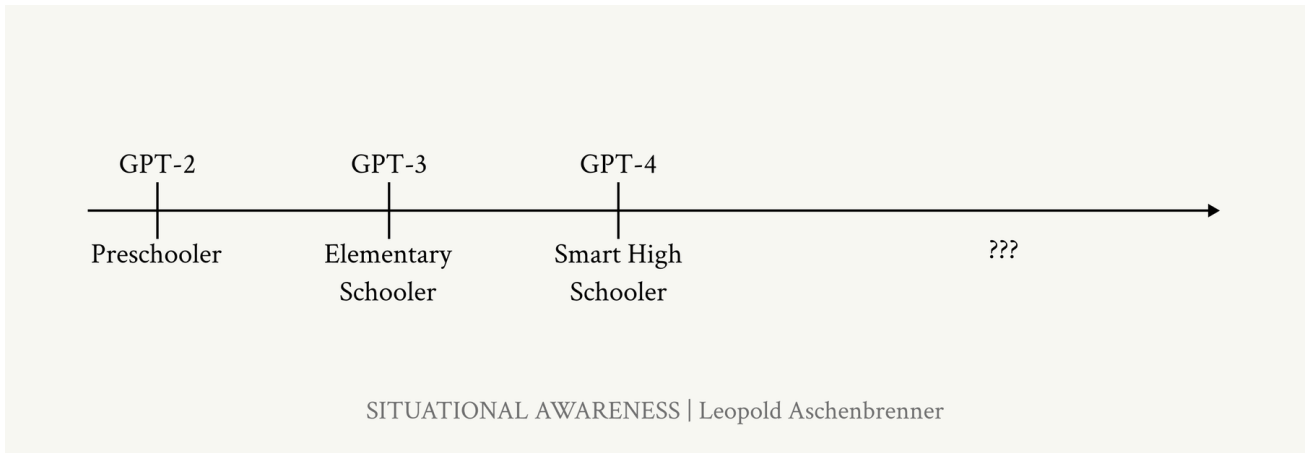


그림 5:단 4년 만의 진전. 당신은 이 선상 어디에 있습니까?

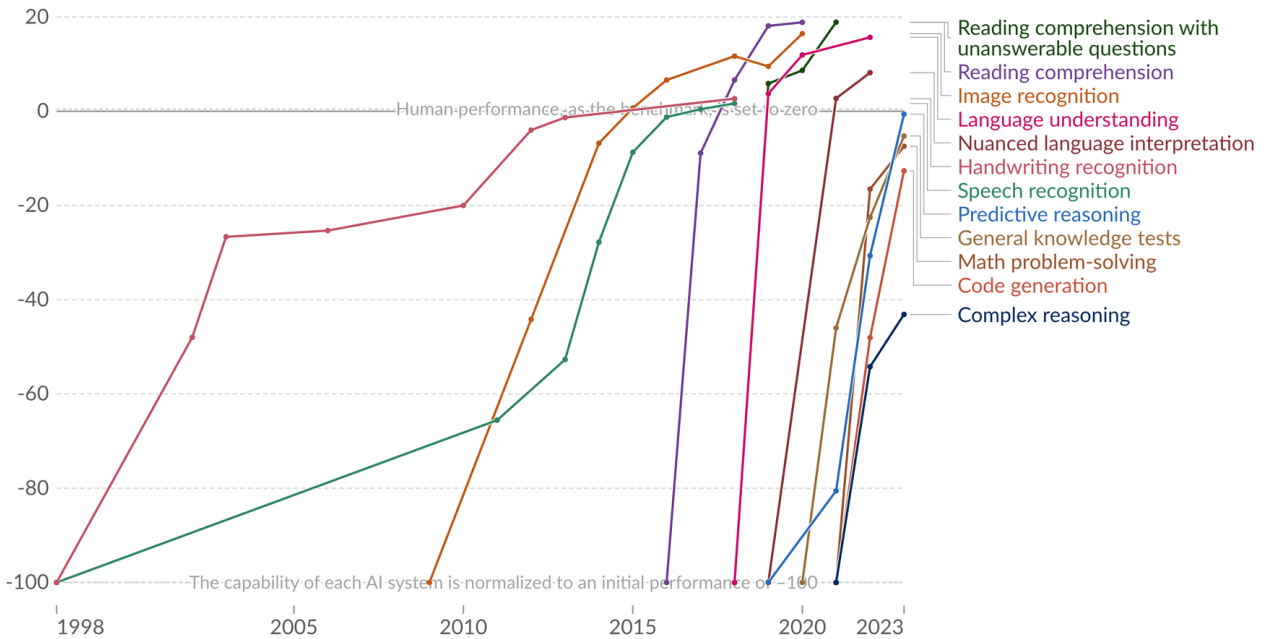
딥러닝의 동향

지난 10년간 딥러닝 진전 속도는 그야말로 놀라웠다. 불과 10년 전만 해도 딥러닝 시스템이 단순한 이미지를 식별하는 것은 혁명적인 일이었다. 오늘날 우리는 계속해서 새롭고 점점 더 어려운 테스트를 고안하려 하지만, 각 새로운 벤치마크는 빠르게 해결된다. 널리 사용되는 벤치마크를 깨는 데 수십 년이 걸리던 시절과 달리, 이제는 몇 달 만에 해결되는 느낌이다.

우리는 문자 그대로 벤치마크가 바닥나고 있다. 일화로, 내 친구들인 댄과 콜린이 2020년에 MMLU라는 벤치마크를 만들었다. 그들은 마침내 고등학생과 대학생에게 주어지는 가장 어려운 시험과 동등한, 시간이 지나도 유효한 벤치마크를 만들고자 했다. 단 3년 만에, GPT-4와 제미니 같은 모델들이 사실상 이를 해결했다.

Test scores of AI systems on various capabilities relative to human performance

Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Data source: Kiela et al. (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.

그림 6: 딥러닝 시스템은 많은 영역에서 인간 수준에 빠르게 도달하거나 이를 능가하고 있다. 그래픽: Our World in Data

약 90%를 달성한다.

더 넓게 보면, GPT-4는 대부분의 표준 고등학교 및 대학 적성 시험을 통과한다(그림 7).⁵

5. 그리고 아니요, 이 테스트들은 훈련 세트에 포함되어 있지 않습니다. AI 연구소들은 좋은 과학을 하기 위해 정확한 측정이 필요하므로 이러한 평가가 오염되지 않도록 진짜 노력을 기울이고 있습니다. 최근 ScaleAI의 분석에 따르면 주요 연구소들은 벤치마크에 과적합하고 있지 않음을 확인했으나(일부 소규모 LLM 개발자들은 수치를 조작할 수 있습니다).

Performance on common exams
(percentile compared to human test-takers)

	GPT-4 (2023)	GPT-3.5 (2022)
Uniform Bar Exam	90th	10th
LSAT	88th	40th
SAT	97th	87th
GRE (Verbal)	99th	63rd
GRE (Quantitative)	80th	25th
US Biology Olympiad	99th	32nd
AP Calculus BC	51st	3rd
AP Chemistry	80th	34th
AP Macroeconomics	92nd	40th
AP Statistics	92nd	51st

SITUATIONAL AWARENESS | Leopold Aschenbrenner

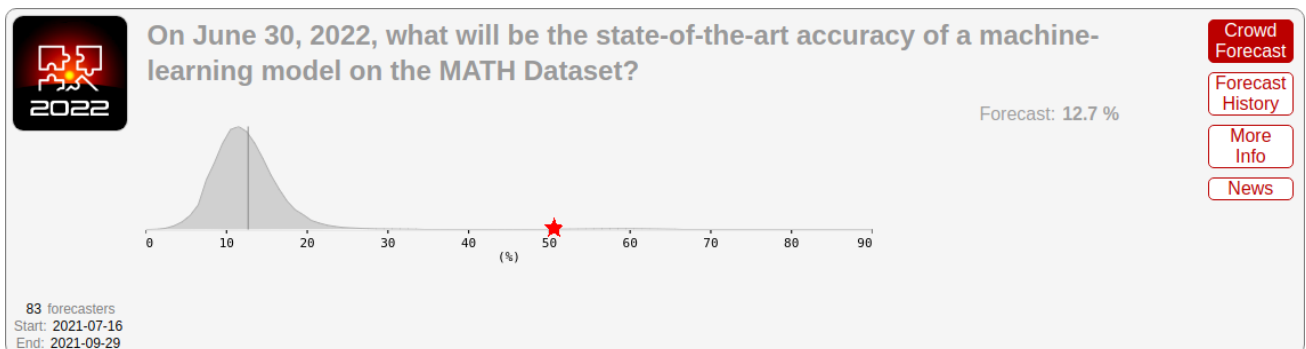


그림 7:표준화된 시험에서의 GPT-4 점수. 또한 GPT-3.5에서 GPT-4로의 인간 백분위 점수의 큰 도약에 주목하라. 이 점수는 종종 평균 이하에서 인간 범위의 최상위권으로 뛰어오른다. (그리고 이것은 GPT-4보다 1년도 채 안 되어 출시된 비교적 최신 모델인 GPT-3.5이며, 앞서 언급한 초등학교 수준의 구식 GPT-3가 아니다!)

Chemistry (general)
<p>A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts?</p> <p>A) A metal compound from the fifth period. B) A metal compound from the fifth period and a non-metal compound from the third period. C) A metal compound from the fourth period. D) A metal compound from the fourth period and a non-metal compound from the second period.</p>
Organic Chemistry
<p>Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)?</p> <p>A) 2 B) 16 C) 8 D) 4</p>
Genetics
<p>If a sperm from species A is injected into an egg from species B and both species have the same number of chromosomes, what would be the main cause of the resulting zygote mortality?</p> <p>A) Species specific zona pellucida proteins on the egg cannot bind sperms from a different species. B) Epistatic interactions between the genes of different species C) Chromosomal incompatibilities will cause failure of meiosis leading to death of zygote. D) Chromosomal recombination will not occur in different species.</p>
Molecular Biology
<p>A scientist studies the stress response of barley to increased temperatures and finds a protein which contributes to heat tolerance through the stabilisation of cell membrane. The scientist is very happy and wants to create a heat-tolerant cultivar of diploid wheat. Using databases, they find a heat tolerance protein homologue and start analysing its accumulation under heat stress. Soon enough, the scientist discovers this protein is not synthesised in the wheat cultivar they study. There are many possible reasons for such behaviour, including:</p> <p>A) A miRNA targets the protein, which makes exonucleases cut it immediately after the end of translation and before processing in ER B) Trimethylation of lysine of H3 histone in position 27 at the promoter of the gene encoding the target protein C) A stop-codon occurs in the 5'-UTR region of the gene encoding the target protein D) The proteolysis process disrupts a quaternary structure of the protein, preserving only a tertiary structure</p>
Astrophysics
<p>Astronomers are studying a star with a Teff of approximately 6000 K. They are interested in spectroscopically determining the surface gravity of the star using spectral lines (EW < 100 mÅ) of two chemical elements, E11 and E12. Given the atmospheric temperature of the star, E11 is mostly in the neutral phase, while E12 is mostly ionized. Which lines are the most sensitive to surface gravity for the astronomers to consider?</p> <p>A) E12 I (neutral) B) E11 II (singly ionized) C) E12 II (singly ionized) D) E11 I (neutral)</p>
Quantum Mechanics
<p>Suppose we have a depolarizing channel operation given by $E(\rho)$. The probability, p, of the depolarization state represents the strength of the noise. If the Kraus operators of the given state are $A_0 = \sqrt{1 - \frac{3p}{4}}$, $A_1 = \sqrt{\frac{p}{4}}X$, $A_2 = \sqrt{\frac{p}{4}}Y$, and $A_3 = \sqrt{\frac{p}{4}}Z$. What could be the correct Kraus Representation of the state $E(\rho)$?</p> <p>A) $E(\rho) = (1 - p)\rho + \frac{p}{3}X\rho X + \frac{p}{3}Y\rho Y + \frac{p}{3}Z\rho Z$ B) $E(\rho) = (1 - p)\rho + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$ C) $E(\rho) = (1 - p)\rho + \frac{p}{4}X\rho X + \frac{p}{4}Y\rho Y + \frac{p}{4}Z\rho Z$ D) $E(\rho) = (1 - p)\rho^2 + \frac{p}{3}X\rho^2 X + \frac{p}{3}Y\rho^2 Y + \frac{p}{3}Z\rho^2 Z$</p>

Table 1: Six example questions from the dataset, two each from subdomains of chemistry, biology, and physics (respectively).

그림 8: 회색: 2021년 8월에 이루어진 전문가 예측으로, 2022년 6월 MATH 벤치마크(고등학교 수학 경시대회)의 어려운 수학 문제) 성능에 대한 예측이다. 빨간 별: 2022년 6월 실제 최첨단 성능으로, 전문가들이 제시한 상한 범위를 훨씬 초과했다. 중간값 ML 연구자들은 더욱 비관적이었다.

7

해마다 반복해서, 회의론자들은 “딥러닝은 X를 할 수 없다”고 주장했지만, 곧바로 틀렸음이 입증되었습니다.⁸ 지난 10년간 AI에서 배운 한 가지 교훈이 있다면, 딥러닝에 절대 베팅하지 말라는 것입니다.

7. 공동 저자는 다음과 같이 언급합니다: “우리 그룹이 처음 MATH 데이터셋을 공개했을 때, 최소한 한 명의 머신러닝 연구자 동료는 ML 모델들이 달성할 수 있는 범위를 크게 벗어나 무의미한 데이터셋이라고 말했습니다(사실 저도 다소 걱정했습니다).”

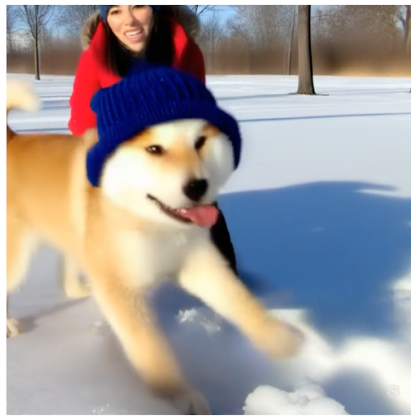
현재 가장 어려운 미해결 벤치마크는 GPQA와 같은 테스트로, 박사 수준의 생물학, 화학, 물리학 문제들로 구성되어 있습니다. 많은 문제들이 저에게는 난해하게 느껴지며, 다른 과학 분야 박사들도 구글을 30분 이상 사용해도 겨우 무작위 확률을 약간 넘는 점수를 받습니다. Claude 3 Opus는 현재 약 60%의 점수를 받고⁹, 해당 분야 박사들은 약 80%를 받는데, 저는 이 벤치마크도 다음 세대나 그 다음 세대에서 무너질 것으로 예상합니다.

여기 2022년에 Yann LeCun이 GPT-5000조차도 실제 세계와의 물리적 상호작용에 대해 추론할 수 없을 것이라고 예측한 내용이 있습니다; GPT-4는 1년 후에 이를 쉽게 수행하는 것이 명백합니다. 여기 GPT-2 이후 해결될 것이라 예측했던 Gary Marcus의 벽과 GPT-3 이후 해결될 것이라 예측했던 벽이 있습니다. 여기 Prof. Bryan Caplan이 그의 첫 공개 내기를 잃은 사례가 있습니다(이전에는 완벽한 공개 내기 기록을 가지고 있었습니다). 2023년 1월, GPT-3.5가 그의 경제학 중간고사에서 D를 받자, Caplan 교수는 2029년까지 어떤 AI도 그의 경제학 중간고사에서 A를 받지 못할 것이라고 Matthew Barnett와 내기를 했습니다. 단 두 달 후, GPT-4가 출시되었고, 즉시 그의 중간고사에서 A를 받았습니다(그는 반에서 가장 높은 점수 중 하나였을 것입니다).

다이아몬드 세트에서, 모델이 사고 과정을 거쳐 32번 시도한 결과에 대한 다수결 투표입니다.



Base compute



4x compute



32x compute

그림 9:GPQA 질문 예시. 모델들은 이미 나보다 이 부분에서 더 뛰어나며, 곧 전문가-박사 수준도 아마 달성할 것이다...

규모 차수(OOM) 세기

이것은 어떻게 일어난 것일까? 딥러닝의 마법은 그저 작동한다는 점이며—반대하는 이들이 끊임없이 있었음에도 불구하고, 추세선은 놀라울 정도로 일관적이었다.

8. 2022년 Yann LeCun이 GPT-5000조차 실제 세계와의 물리적 상호작용을 추론할 수 없을 것이라고 예측했으나, 1년 후 GPT-4는 명백히 이를 쉽게 수행합니다. GPT-2 이후 해결될 것이라 예측했던 Gary Marcus의 한계와 GPT-3 이후 해결될 것이라 예측했던 한계를 GPT-4가 극복했습니다. Bryan Caplan 교수는 그의 완벽한 공적 내기 기록 후 첫 공개 내기에서 패배했습니다. 2023년 1월, GPT-3.5가 그의 경제학 중간고사에서 D를 받자 Caplan 교수는 2029년까지 AI가 그의 경제학 중간고사에서 A를 받을 수 없을 것이라고 Matthew Barnett과 내기했습니다. 단 두 달 후 GPT-4가 출시되자 즉시 그의 중간고사에서 A를 받았으며, 이는 그의 반에서 가장 높은 점수 중 하나였습니다.

9. 다이아몬드 세트에서는, 모델이 사슬 추론(chain-of-thought)으로 32회 시도한 결과를 다수결로 결정했습니다.

각 규모 차수(OOM)의 유효 컴퓨팅이 증가할 때마다, 모델은 예측 가능하고 신뢰할 수 있게 더 나아진다.¹⁰ 만약 우리가 OOM을 셀 수 있다면, (대략적으로, 정성적으로) 능력 향상을 외삽할 수 있다.¹¹ 이것이 몇몇 선견지명이 있는 이들이 GPT-4의 등장을 예측한 방법이다.

우리는 GPT-2에서 GPT-4까지 4년간의 진보를 세 가지 규모 확장 범주로 분해할 수 있다:

1. 컴퓨팅: 우리는 이 모델들을 훈련시키기 위해 훨씬 더 큰 컴퓨터를 사용하고 있다.
2. 알고리즘 효율성: 알고리즘 진보의 지속적인 추세가 있다. 이들 중 다수는 “컴퓨팅 승수” 역할을 하며, 우리는 이를 증가하는 유효 컴퓨팅의 통합 척도로 놓을 수 있다.
3. 제약 해소(unhobbling) 이득: 기본적으로 모델은 많은 놀라운 원시 능력을 학습하지만, 여러 어리석은 방식으로 제약되어 실제 가치가 제한된다. 인간 피드백 강화 학습(RLHF), 사고의 연쇄(CoT), 도구, 그리고 스캐폴딩과 같은 간단한 알고리즘 개선을 통해 상당한 잠재 능력을 해방할 수 있다.

Training compute of notable models

EPOCH AI

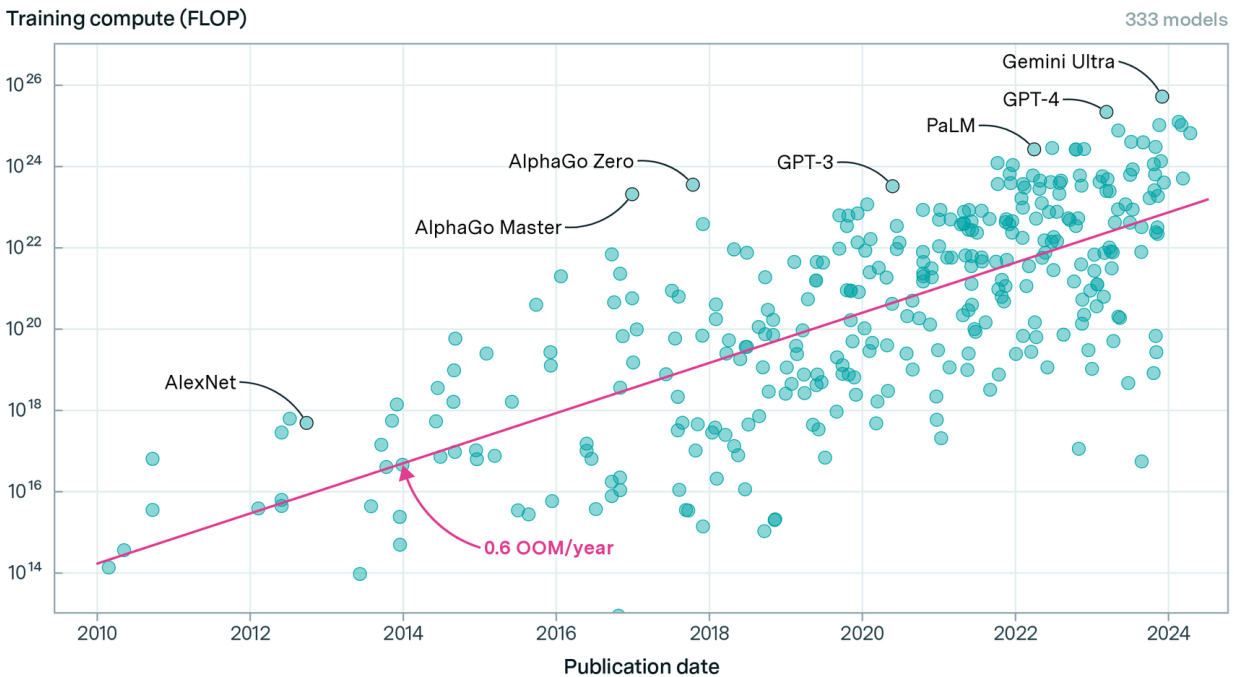


그림 10: OpenAI Sora의 예에서 컴퓨팅 규모 확장의 효과.

10. 그리고 이 추세선이 얼마나 일관된지 주목할 만합니다. 원래의 스케일링 법칙 논문과 이후의 연산량 및 연산 효율 스케일링 추정치를 결합하면, 15자리수 이상의 규모(1,000조 배 이상의 실효 컴퓨트 규모)에 걸쳐 일관된 스케일링 추세를 시사합니다!

11. 흔한 오해는 스케일링이 당황도 손실(perplexity loss)에만 해당한다는 것이지만, 우리는 벤치마크 하위 과제의 성능에서도 매우 명확하고 일관된 스케일링 거동을 관찰합니다. 보통 올바른 로그-로그 그래프를 찾는 문제일 뿐입니다. 예를 들어 GPT-4 블로그 포스트에서는 MLPR(평균 로그 통과율)를 사용하여 약 6 OOM(1,000,000배)의 컴퓨트 스케일에 걸쳐 코딩 문제 성능에서 일관된 스케일링 거동을 보여줍니다. “Are Emergent Abilities a Mirage?” 논문도 유사한 요점을 제시합니다; 올바른 지표 선택 시 하위 과제 성능에 대해 거의 항상 일관된 추세가 있습니다. 더 일반적으로, “스케일링 가설”이라는 정성적 관찰—스케일에 따른 모델 역량의 매우 명확한 추세—은 손실 스케일링 곡선 이전부터 존재했으며, “스케일링 법칙” 작업은 이를 정식으로 측정하는 것입니다.

그리고 이 추세선들이 얼마나 일관적인지 주목할 가치가 있다. 원래의 규모 확장 법칙 논문과 이후 컴퓨팅 및 컴퓨팅 효율성 확장에 대한 일부 추정치를 결합하면, 15개 이상의 규모 차수(1,000조 배 이상의 유효 컴퓨팅)에 걸쳐 일관된 확장 추세가 있음을 시사한다!

흔한 오해 중 하나는 규모 확장이 오직 당혹도 손실(perplexity loss)에만 적용된다는 것이다. 하지만 우리는 벤치마크에서 하위 작업 성능에 대해서도 매우 명확하고 일관된 확장 행동을 관찰한다. 보통은 적절한 로그-로그 그래프를 찾는 문제일 뿐이다. 예를 들어, GPT-4 블로그 게시물에서는 MLPR(평균 로그 통과율)을 사용해 6 OOM(100만 배)의 컴퓨팅에 걸쳐 코딩 문제 성능에 대한 일관된 확장 행동을 보여준다.

“Are Emergent Abilities a Mirage?” 논문도 비슷한 점을 지적한다; 적절한 지표 선택 시 하위 작업 성능에 대해 거의 항상 일관된 추세가 있다.

더 일반적으로, “규모 확장 가설”이라는 정성적 관찰—규모에 따른 모델 능력의 매우 명확한 추세—은 손실-확장 곡선 이전부터 존재했으며, “규모 확장 법칙” 연구는 이를 공식적으로 측정한 것이다.

우리는 이러한 축을 따라 개선의 규모 차수(OOM)를 “셈”할 수 있습니다: 즉, 각 축에 대해 유효 컴퓨팅 단위로 확장 규모를 추적하는 것입니다. 3배는 0.5 OOM; 10배는 1 OOM; 30배는 1.5 OOM; 100배는 2 OOM; 이런 식입니다. 또한 2023년부터 2027년까지 GPT-4 이후에 무엇을 기대할 수 있을지도 살펴볼 수 있습니다.

각 항목을 하나씩 살펴보겠지만, 결론은 명확합니다: 우리는 OOM을 빠르게 돌파하고 있습니다. 데이터 벽에서 잠재적 역풍이 있을 수 있는데, 이에 대해서도 다루겠지만—전반적으로 2027년까지 GPT-4 위에 GPT-2에서 GPT-4로의 점프와 비슷한 규모의 도약을 기대하는 것이 타당해 보입니다.

컴퓨팅

최근 진보의 가장 일반적으로 논의되는 동인부터 시작하겠습니다: 모델에 (엄청난) 컴퓨팅을 투입하는 것입니다.

많은 사람들은 이것이 단순히 무어의 법칙 때문이라고 생각합니다. 하지만 무어의 법칙이 전성기였던 옛날에도 그것은 비교적 느린 속도였으며—아마도 10년에 1~1.5 OOM 정도였습니다. 우리는 훨씬 더 빠른 컴퓨팅 확장을 보고 있는데—무어의 법칙 속도의 거의 5배에 달하는 속도—이는 거대한 투자 덕분입니다. (한 모델에 백만 달러를 쓰는 것은 예전에는 상상도 못할 일이었지만, 이제는 푼돈에 불과합니다!)

표 1: Epoch AI의 GPT-2부터 GPT-4까지 컴퓨팅 추정치 모델 추정 컴퓨팅량 성장 GPT-2 (2019) 약 $4e^{21}$ FLOP GPT-3 (2020) 약 $3e^{23}$ FLOP + 약 2 OOM GPT-4 (2023) $8e^{24} \sim 4e^{25}$ FLOP + 약 1.5~2 OOM

우리는 AI 트렌드에 대한 탁월한 분석으로 널리 존경받는 Epoch AI의 공개 추정치를 사용해 2019년부터 2023년까지 컴퓨팅 확장 규모를 추적할 수 있습니다. GPT-2에서 GPT-3로의 확장은 빠른 확장이었으며; 작은 실험에서 대규모 언어 모델을 훈련시키기 위해 전체 데이터센터를 사용하는 규모로 확장하는 큰 여유가 있었습니다. GPT-3에서 GPT-4로의 확장에서는 완전히 새로운 (훨씬 더 큰) 클러스터를 구축해야 하는 현대적 체제로 전환했습니다. 그리고

극적인 성장은 계속되었다. 전반적으로 Epoch AI의 추정에 따르면 GPT-4 훈련에는 GPT-2보다 약 3,000배에서 10,000배 더 많은 원시 컴퓨팅이 사용되었다.

대략적으로 보면, 이는 단지 더 오래 지속된 추세의 연장선일 뿐이다. 지난 15년 동안, 주로 투자 규모의 대규모 확대(그리고 GPU 및 TPU 형태의 AI 작업 부하에 특화된 칩 개발) 덕분에, 최첨단 AI 시스템에 사용되는 훈련 컴퓨팅은 연간 약 0.5 규모 차수(OOM)씩 증가해왔다.

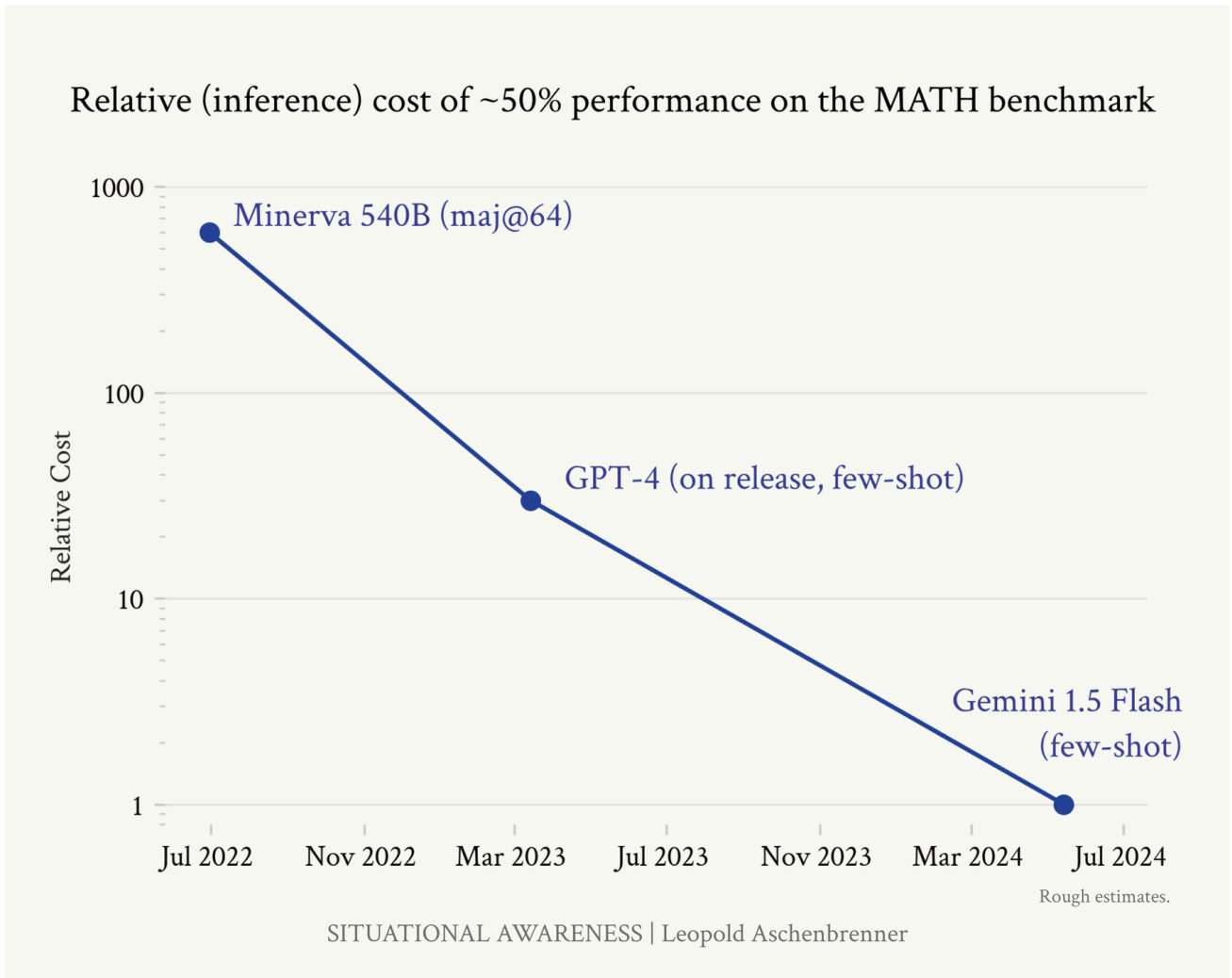


그림 11: 시간에 따른 주목할 만한 딥러닝 모델들의 훈련 컴퓨팅. 출처: Epoch AI.

GPT-2에서 GPT-3로의 1년간 컴퓨팅 규모 확대는 이례적인 과도기였지만, 모든 징후는 장기 추세가 계속될 것임을 시사한다. 샌프란시스코 루머 시장은 대규모 GPU 주문에 관한 극적인 이야기로 가득하다. 관련 투자는 엄청날 것이지만 이미 진행 중이다. 이 점에 대해서는 시리즈의 IIIa. ‘조 단위 클러스터 경쟁’에서 더 자세히 다룰 것이다; 그 분석에 따르면, 2027년 말까지 추가로 2 규모 차수(OOM)의 컴퓨팅(수십억 달러 규모의 클러스터)이 발생할 가능성이 매우 높으며, 3 규모 차수 이상(1,000억 달러 이상) 클러스터도 가능성이 있으며(마이크로소프트/오픈AI에서 진행 중이라는 루머가 있다).

알고리즘 효율성

컴퓨팅에 대한 막대한 투자가 모든 주목을 받는 반면, 알고리즘 진보도 아마 비슷하게 중요한 진보의 원동력이며(그리고 극적으로 과소평가되어 왔다).

알고리즘 진보가 얼마나 큰 영향을 미칠 수 있는지 보기 위해, 단 2년 만에 MATH 벤치마크(고등학교 수학 경시대회 수학)에서 약 50% 정확도를 달성하는 데 드는 비용이 얼마나 떨어졌는지에 대한 다음 그림(그림 12)을 고

려해 보자. (비교를 위해, 수학을 별로 좋아하지 않는 컴퓨터 과학 박사과정 학생이 40%를 기록했으니, 이미 꽤 좋은 수준이다.) 추론 효율성은 2년도 채 안 되어 거의 3 규모 차수(OOM)—1,000배—개선되었다.

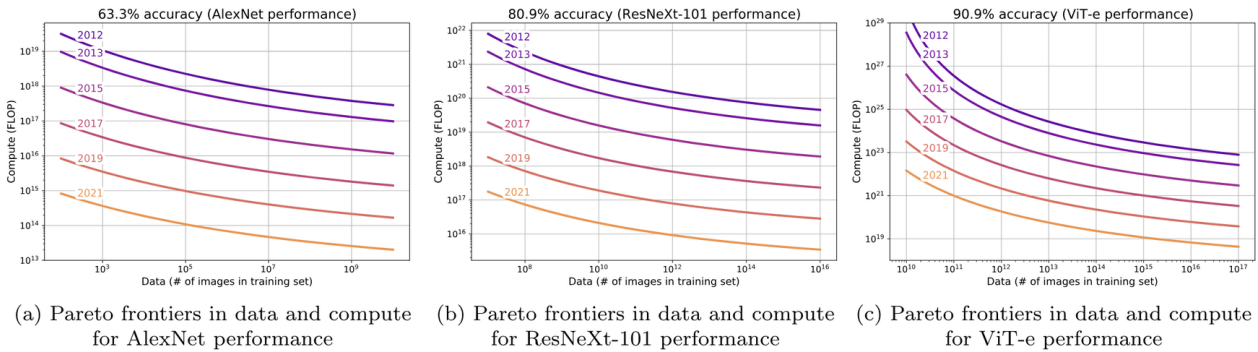


Figure 1. Pareto frontiers for training models to achieve performance of well-known models over time.

그림 12: 약 50% MATH 성능 달성에 필요한 상대적 추론 비용에 대한 대략적 추정. 아래 계산 참조. Gemini 1.5 Flash는 MATH에서 54.9%를 기록하며, 백만 토큰당 \$0.35/\$1.05(입력/출력) 비용이 든다. GPT-4는 사전 공개 시 MATH에서 42.5%, 2023년 초에는 52.9%를 기록했고, 백만 토큰당 \$30/\$60(입력/출력) 비용이 든다; 이는 Gemini 1.5 Flash보다 토큰당 85배/57배(입력/출력) 더 비싸다. 보수적으로, Gemini 1.5 Flash가 문제를 해결하는 데 더 많은 토큰을 사용할 가능성을 감안해 30배 비용 감소 추정을 사용한다. Minerva540B는 MATH에서 50.3%를 기록하며, 64개 샘플의 다수결 투표를 사용한다. 지식이 있는 친구는 여기서 기본 모델이 GPT-4보다 추론 비용이 아마 2~3배 더 비쌀 것으로 추정한다. 그러나 Minerva는 빠른 점검에서 답변당 사용하는 토큰 수가 다소 적은 것으로 보인다. 더 중요한 점은, Minerva가 그 성능을 달성하기 위해 64개의 샘플이 필요했으며, 이는 예를 들어 추론 API를 통해 단순히 실행한다면 비용이 64배 증가함을 의미한다. 실제로는 평가를 실행할 때 프롬프트 토큰을 캐시할 수 있다; 몇 샷 프롬프트를 고려하면, 출력 토큰을 감안해도 프롬프트 토큰이 비용의 대부분일 가능성이 크다. 출력 토큰이 단일 샘플 비용의 3분의 1이라고 가정하면, 캐시를 고려한 다수결 64회의 비용 증가는 약 20배에 불과하다. 보수적으로, 위에서 20배 비용 감소라는 대략적 수치를 사용한다 (API를 통해 실행할 경우 추론 비용 감소는 더 클 수 있지만).

이 수치들은 추론 효율성에 관한 것일 뿐(이는 훈련 효율성 개선과 일치하지 않을 수도 있으며, 훈련 효율성은 공개 데이터에서 추론하기 더 어렵다)임에도 불구하고, 엄청난 양의 알고리즘 진보가 가능하며 실제로 일어나고 있음을 명확히 보여준다.

이 글에서는 두 가지 종류의 알고리즘 진보를 구분하겠다. 여기서는 먼저 “패러다임 내” 알고리즘 개선, 즉 단순히 더 나은 기본 모델을 만들어내고, 컴퓨팅 효율성 또는 컴퓨팅 배수로서 직접 작용하는 개선부터 다룬다. 예를 들어, 더 나은 알고리즘은 동일한 성능을 10배 적은 훈련 컴퓨팅으로 달성할 수 있게 해준다. 이는 결과적으로 유효 컴퓨팅이 10배(1 규모 차수, OOM) 증가하는 효과를 낸다. (나중에 “제약 해소”를 다룰 텐데, 이는 기본 모델의 능력을 확장하는 “패러다임 확장/응용 확장” 알고리즘 진보로 생각할 수 있다.)

장기적 추세를 살펴보면, 새로운 알고리즘 개선은 꽤 일정한 속도로 발견되는 것 같다. 개별 발견은 무작위적이고, 매번 극복하기 어려운 장애물이 존재하는 것처럼 보이지만, 장기 추세선은 예측 가능하며 그래프상 직선 형태를 띤다. 추세선을 신뢰하라.

가장 좋은 데이터는 ImageNet에 대해 가지고 있다(여기서는 알고리즘 연구가 대부분 공개되어 있고, 10년 이상에 걸친 데이터가 존재한다). 2012년부터 2021년까지 9년 동안 컴퓨팅 효율성이 연평균 약 0.5 규모 차수(OOM)씩 꾸준히 개선되었다.

그것은 엄청난 의미를 가진다: 4년 후에는 동일한 성능을 약 100배 적은 컴퓨팅으로 달성할 수 있다는 뜻이며(동시에, 동일한 컴퓨팅으로 훨씬 더 높은 성능을 낼 수 있다!).

불행히도 연구소들이 내부 데이터를 공개하지 않기 때문에, 최첨단 대형 언어 모델(LLM)에 대한 알고리즘 진보를 측정하기는 더 어렵다.

지난 4년 동안, EpochAI는 ImageNet에서의 결과를 언어 모델링에 복제하는 새로운 작업을 진행했으며, 2012년부터 2023년까지 LLM에서 약 0.5 규모 차수(OOMs)/년의 알고리즘 효율성 추세를 추정한다. (하지만 이 추정치는 오차 범위가 더 넓고, 주요 연구소들이 알고리즘 효율성 발표를 중단했기 때문에 최근의 일부 향상은 반영하지 못한다.)

Efficiency doubles roughly every 8 months

Effective compute (relative to 2014)

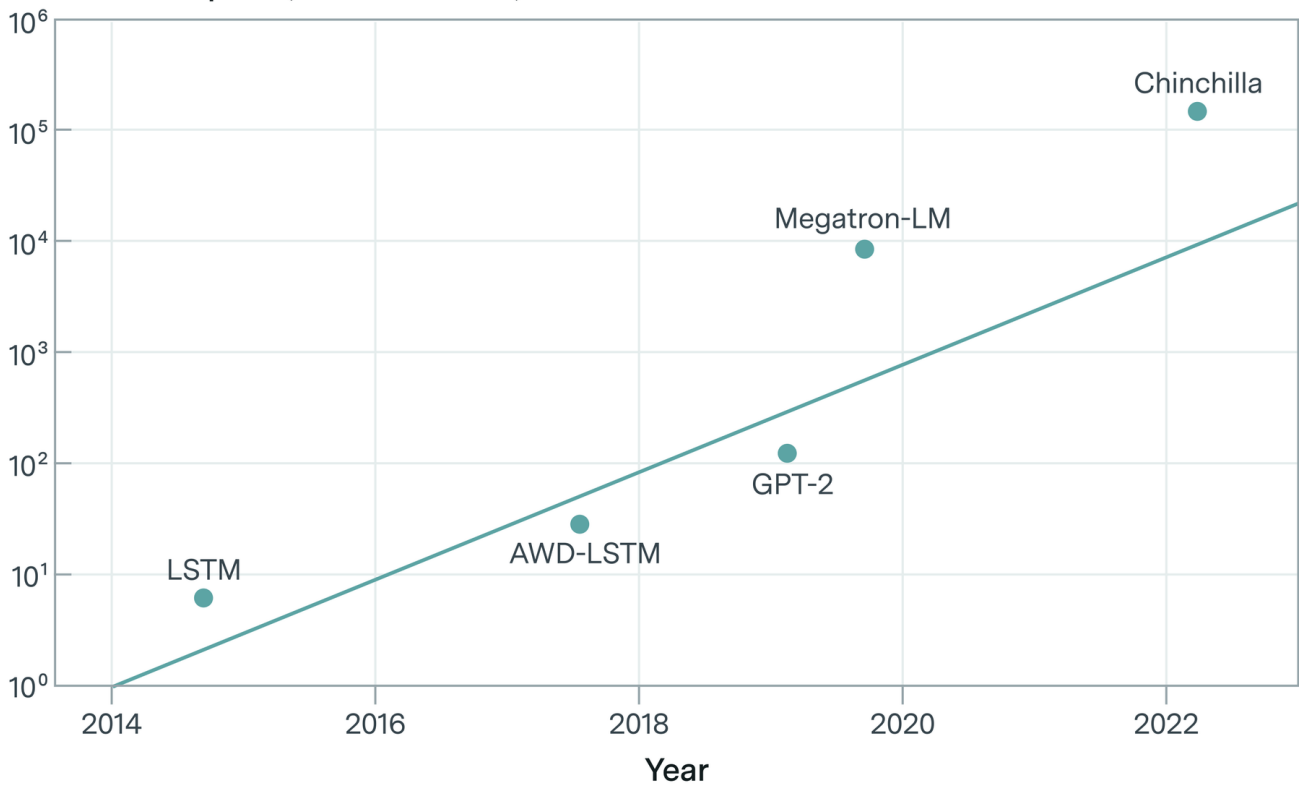


그림 14: Epoch AI가 추정한 언어 모델링에서의 알고리즘 효율성. 이들의 추정에 따르면 8년 동안 약 4 규모 차수의 효율성 향상이 이루어졌다.

이들은 추론 효율성(반드시 훈련 효율성은 아님)이며, 어느 정도는 추론 특화 최적화를 반영하지만, a) 이는 엄청난 양의 알고리즘 진보가 가능하며 실제로 일어나고 있음을 시사하고, b) 알고리즘 개선은 종종 훈련 효율성과 추론 효율성 모두에 해당하는 경우가 많다. 예를 들어, 지난 4년을 좀 더 직접적으로 보면, GPT-2에서 GPT-3로의 전환은 기본적으로 단순한 규모 확장(논문에 따르면)이었지만, GPT-3 이후로 공개적으로 알려지고 추론 가능한 많은 향상이 있었다: 파라미터 수를 줄임으로써 가능하다. • API 비용에서 향상을 추론할 수 있다:¹² - GPT-4는 출시 당시 GPT-3가 출시되었을 때와 거의 같은 비용이었음에도 불구하고 성능은 엄청나게 향상되었다.¹³ (규모 법칙에 기반한 단순하고 대략적인 추정에 따르면, GPT-3에서 GPT-4로의 유효 컴퓨팅 증가의 약 절

12. 이들은 추론 효율(inference efficiencies)이긴 하지만(훈련 효율로 반드시 연결되지는 않음) 어느 정도 추론 특화 최적화를 반영할 수 있으며, a) 대규모 알고리즘적 진전이 가능하고 실제로 진행 중임을 시사하며, b) 알고리즘 개선은 통상적으로 파라미터 수를 줄임으로써 훈련 효율과 추론 효율 모두를 개선하는 경우가 많습니다.

반이 알고리즘 개선에서 비롯된 것으로 보인다.¹⁴⁾ - GPT-4 출시 1년 후, GPT-4o 출시와 함께 OpenAI의 GPT-4 수준 모델 가격은 입력/출력 기준으로 각각 6배/4배 더 하락했다.

GPT-3: 100만 토큰당 \$60, GPT-4: 100만 입력 토큰당 \$30, 100만 출력 토큰당 \$60. Chinchilla 규모 법칙에 따르면 파라미터 수와 데이터 양을 동일하게 확장해야 한다. 즉, 파라미터 수는 유효 훈련 컴퓨팅이 증가하는 규모 차수의 절반만큼 증가한다. 동시에 파라미터 수는 직관적으로 추론 비용과 대략 비례한다. 다른 조건이 같다면, 추론 비용이 일정하다는 것은 유효 컴퓨팅 증가의 규모 차수 절반이 알고리즘 개선으로 “상쇄”되었음을 의미한다. 다만, 명확히 하자면, 이는 매우 단순한 계산(대략적인 설명용)으로, 여러 면에서 부정확하다. 추론 특화 최적화(훈련 효율성으로 이어지지 않는)가 있을 수 있고, 파라미터 수를 줄이지 않는 훈련 효율성(추론 효율성으로 이어지지 않는)도 있을 수 있으며, 기타 여러 요인이 있다.

- 최근 출시된 Gemini 1.5 Flash는 “GPT-3.75 수준”과 GPT-4 수준 사이의 성능을 제공하며,¹⁵ 원래 GPT-4 대비 입력/출력 기준으로 85배/57배 적은 비용으로(엄청난 향상!) 작동한다.

종합하면, 공개된 정보는 GPT-2에서 GPT-4로의 도약에 1~2 규모 차수의 알고리즘 효율성 향상이 포함되었음을 시사한다.¹⁷¹⁸

13. GPT-3: 100만 토큰당 60달러, GPT-4: 100만 입력 토큰당 30달러, 100만 출력 토큰당 60달러.

14. Chinchilla 스케일링 법칙에 따르면 파라미터 수와 데이터 양을 동일하게 스케일해야 합니다. 즉, 파라미터 수는 실효 훈련 컴퓨트 증가의 ‘절반 OOM’만큼 성장합니다. 동시에, 파라미터 수는 직관적으로 대략 추론 비용과 비례합니다. 모든 것이 동일하다면, 일정한 추론 비용은 실효 컴퓨트 증가의 절반 OOM이 알고리즘 개선 효과에 의해 ‘상쇄’되었음을 의미합니다. 그럼에도 이것은 매우 단순한 계산일 뿐이며(대략적 설명용) 여러 면에서 부정확할 수 있습니다. 추론 특화 최적화가 있을 수 있고(훈련 효율로 연결되지 않을 수도 있음), 훈련 효율이 증가해도 파라미터 수가 줄지 않아 추론 효율로 연결되지 않는 경우도 있습니다.

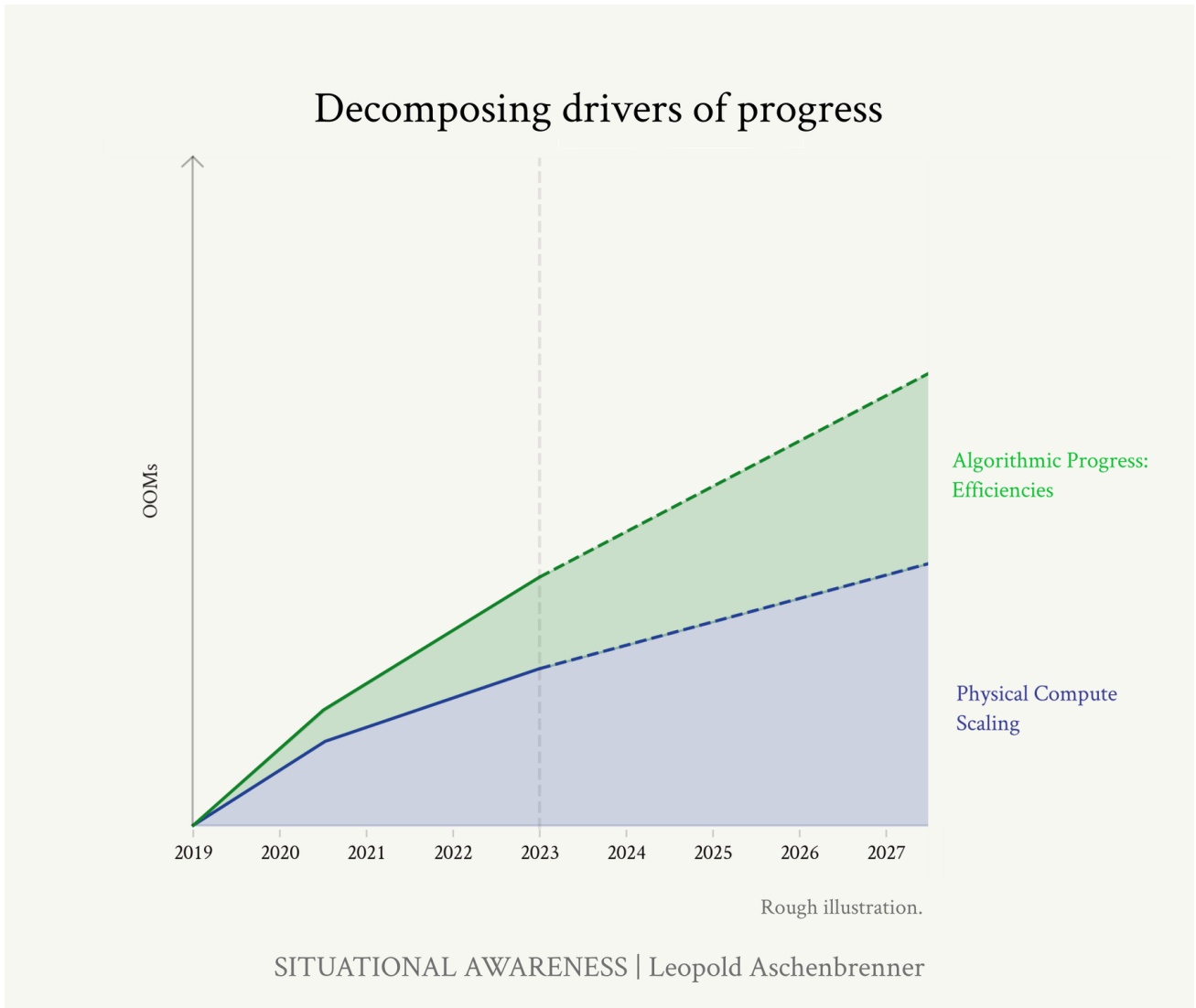


그림 15:진보 분해: 컴퓨팅과 알고리즘 효율성. (대략적인 도식.)

GPT-4 이후 4년 동안, 이 추세는 계속될 것으로 예상된다.¹⁹ 평균적으로 연간 약 0.5 규모 차수(OOM)의 컴퓨팅 효율성 향상, 즉 2027년까지 GPT-4 대비 약 2 OOM의 이득이 있을 것이다. 컴퓨팅 효율성은 저절로 얻기 쉬운 성과가 줄어들면서 찾기 어려워지겠지만, AI 연구소들은 새로운 알고리즘 개선을 찾기 위한 자금과 인재 투자를 빠르게 늘리고 있다.²⁰ (공개적으로 추론 가능한 추론 비용 효율성은 적어도 전혀 둔화되지 않은 것으로 보인다.) 최상위에서는, 트랜스포머와 같은²¹ 근본적인 돌파구가 더 큰 이득을 가져올 수도 있다.

종합하면, 2027년 말까지 GPT-4 대비 약 1~3 OOM의 알고리즘 효율성 향상을 기대할 수 있으며, 최선의 추정치는 약 2 OOM 정도이다.

데이터 장벽

15. Gemini 1.5 Flash는 LMSys 챗봇 리더보드에서 GPT-4와 비슷한 순위(원래 GPT-4보다 높고, 업데이트된 GPT-4보다는 낮음)를 기록하며, MATH와 GPQA(추론 평가)에서는 원래 GPT-4와 비슷한 성능을 보이고, MMLU(지식 평가 중점)에서는 GPT-3.5와 GPT-4의 중간 정도 위치에 있습니다.

17. 예를 들어, 이 논문에서는 GPT-3 스타일의 기본 트랜스포머를 수년간 발표된 아키텍처 및 훈련 레시피의 단순 변경 사항들과 비교합니다 (RMSnorm 대신 layernorm, 다양한 위치 임베딩, SwiGlu 활성화, Adam 대신 AdamW 옵티마이저 등). 이들을 ‘Transformer++’라 부르며 소규모 스케일에서 최소 6배 향상을 시사합니다.

이 모든 것에 대해 잠재적으로 중요한 변동 요인이 있다: 인터넷 데이터가 고갈되고 있다는 점이다. 이는 곧, 더 많은 스크랩된 데이터를 이용해 더 큰 언어 모델을 사전학습하는 단순한 접근법이 심각한 병목에 부딪힐 수 있음을 의미한다.

최첨단 모델들은 이미 인터넷의 대부분을 학습에 사용했다. 예를 들어 Llama 3는 15조 토큰 이상으로 학습되었다. LLM 학습에 사용되는 인터넷 데이터 덤프인 Common Crawl은 원시 데이터 기준 100조 토큰 이상이지만, 그중 상당 부분은 스팸과 중복이다(예를 들어, 비교적 단순한 중복 제거 후 30조 토큰이 남는데, 이는 Llama 3가 사실상 모든 데이터를 이미 사용하고 있음을 시사한다). 게다가 코드와 같은 특정 도메인에서는 토큰 수가 훨씬 적은데, 예를 들어 공개된 깃허브 저장소는 수조 토큰 수준으로 추정된다.

데이터를 반복해서 사용할 수는 있지만, 이에 관한 학술 연구는 반복이 한계가 있음을 보여준다.¹⁶ 예폭(16배 반복) 이후에는 수익이 매우 빠르게 감소하여 거의 없어진다. 어느 시점에서는 더 많은 (유효) 컴퓨팅 자원이 있어도 데이터 제약 때문에 모델을 개선하는 것이 훨씬 어려워질 수 있다. 이는 과소평가해서는 안 된다: 우리는 지금까지 계속해서

스케일링 곡선, 언어 모델링-사전 학습-패러다임의 물결을 타고, 여기서 새로운 무언가가 없으면 이 패러다임은 (적어도 순진하게 보면) 한계에 도달할 것이다. 막대한 투자가 있었음에도 불구하고, 우리는 정체기에 이를 것이다.

모든 연구소들은 이를 극복하기 위한 새로운 알고리즘 개선이나 접근법에 대해 막대한 연구 베팅을 하고 있다는 소문이 돌고 있다. 연구자들은 합성 데이터부터 자기 대결(self-play)과 강화 학습(RL) 접근법에 이르기까지 다양한 전략을 시도하고 있다고 한다. 업계 내부자들은 매우 낙관적인 것으로 보인다. Anthropic의 CEO 다니오 아모데이(Dario Amodei)는 최근 팟캐스트에서 “매우 순진하게 보면 데이터가 바닥나는 것과 그리 멀지 않다 [...] 내 추측으로는 이것이 장애물이 되지 않을 것이다 [...] 이를 해결할 방법은 매우 다양하다”고 말했다. 물론, 이와 관련된 연구 결과들은 현재 비공개이며 발표되지 않고 있다.

내부자의 낙관론 외에도, 훨씬 더 나은 샘플 효율성(제한된 데이터로부터 더 많이 학습할 수 있게 하는 알고리즘 개선)을 통해 모델을 훈련시키는 방법을 찾을 수 있을 것이라는 강한 직관적 근거가 있다고 생각한다. 당신이나 내가 정말 밀도 높은 수학 교과서를 어떻게 학습할지 생각해 보라:

- 현대의 대형 언어 모델(LLM)이 훈련 중 하는 일은 본질적으로 교과서를 매우 빠르게 훑어보는 것이다. 단어들 이 그냥 지나가며, 많은 뇌력을 들이지 않는다.
- 반면, 당신이나 나는 그 수학 교과서를 읽을 때 몇 페이지를 천천히 읽고, 머릿속으로 그 내용을 내적 독백하며 몇몇 스터디 친구들과 토론한다. 그리고 한두 페이지를 더 읽고, 연습 문제를 풀어보고 실패하고, 다른 방식으로 다시 시도하고, 문제에 대한 피드백을 받고, 맞을 때까지 반복한다. 그렇게 하다 보면 결국 내용이 ‘딱’ 이해된다.
- 만약 당신이나 내가 LLM처럼 교과서를 그냥 훑어보는 것밖에 할 수 없다면, 밀도 높은 수학 교과서에서 거의 아무것도 배우지 못할 것이다.²²

18. 0.5 규모 차수/년의 추세와 GPT-2와 GPT-4 출시 사이 4년을 고려하면 2 규모 차수가 된다. 그러나 GPT-2에서 GPT-3는 단순한 규모 확장이었고(예: Transformer로부터 큰 향상 후), OpenAI는 GPT-4 사전학습을 2022년에 완료했다고 주장하므로, 여기서 계산해야 할 알고리즘 진보 기간은 2년에 더 가까울 수 있다. 1 규모 차수의 알고리즘 효율성은 보수적인 하한으로 보인다.

19. 최소한, 10년 넘는 일관된 알고리즘 개선을 감안할 때, 모든 발전이 갑자기 멈출 것이라 주장하는 쪽에 증명 책임이 있습니다!

- 하지만 아마도, 인간이 뻘뻘한 수학 교과서를 소화하는 방식을 일부 도입하여 모델이 제한된 데이터로부터 훨씬 더 많이 학습할 수 있는 방법이 있을 것이다. 23 또 다른 흥미로운 사고 방식은 다음과 같다: 사전 학습과 문맥 내 학습 사이에 “중간 결손(missing-middle)”이 존재한다는 것이다. 단순화해서 말하면, 이런 종류의 것—자료에 대한 내적 독백, 스터디 버디와의 토론, 문제를 시도하고 실패하다가 결국 이해하는 과정—은 많은 합성 데이터/자기 대국/강화 학습(RL) 접근법이 시도하는 바이다.²³

과거 최첨단 모델 훈련 방식은 단순하고 순진했지만 효과가 있었기에, 아무도 표본 효율성을 깨기 위해 열심히 노력하지 않았다. 이제 그것이 제약이 될 수 있으므로, 모든 연구소가 수십억 달러와 최고의 인재를 투입해 이를 해결하려 할 것이다. 딥러닝에서 흔한 패턴은 세부 사항을 맞추기 위해 많은 노력(그리고 실패한 프로젝트들)이 필요하지만, 결국 명백하고 단순한 어떤 버전이 작동한다는 것이다. 지난 10년간 딥러닝이 모든 장벽을 돌파해 온 방식을 고려하면, 여기서도 비슷할 것이라는 것이 내 기본 가정이다.

더욱이, 합성 데이터 같은 알고리즘적 베틱 중 하나를 깨는 것이 모델을 극적으로 향상시킬 가능성도 실제로 있어 보인다. 직관을 돕기 위해 설명하자면, 현재 최첨단 모델인 Llama 3 같은 모델은 인터넷에서 훈련되는데, 인터넷은 대부분 전자상거래나 SEO 같은 쓰레기 콘텐츠다. 많은 대형 언어 모델들은 훈련 컴퓨팅의 대부분을 이런 쓰레기 데이터에 소비하며, 어려운 과학 문제를 해결하는 사람들의 추론 과정 같은 고품질 데이터에는 거의 쓰지 않는다. 만약 GPT-4 수준의 컴퓨팅을 완전히 고품질 데이터에만 쓸 수 있다면, 훨씬 더 뛰어난 모델이 될 수 있다.

알파고를 되돌아보는 것도 유용하다. 알파고는 세계 챔피언을 이긴 최초의 AI 시스템으로, 그 가능성이 논의된 지 수십 년 만에 이루어진 성과다.²⁴

- 1단계에서 알파고는 전문가 인간 바둑 기보를 모방 학습하여 훈련되었다. 이것이 기초를 마련했다.
- 2단계에서 AlphaGo는 스스로 수백만 판의 게임을 플레이했습니다. 이를 통해 인간을 능가하는 바둑 실력을 갖추게 되었죠. 이세돌과의 대국에서 나온 유명한 37수, 인간이라면 절대 두지 않았을 매우 이례적이면서도 탁월한 수를 기억하세요.

LLM에 대해 2단계에 해당하는 과정을 개발하는 것은 데이터 장벽을 극복하기 위한 핵심 연구 과제입니다(더 나아가 궁극적으로 인간 수준의 지능을 뛰어넘는 열쇠가 될 것입니다).

이 모든 것은 데이터 제약이 향후 AI 발전을 예측하는 데 큰 오차 범위를 불러일으키는 것처럼 보인다는 점을 의미합니다. 상황이 정체될 가능성도 매우 현실적입니다(LLM이 인터넷만큼 중요한 존재가 될 수는 있지만, 진정한 의미의 미친 AGI에 도달하지 못할 수도 있습니다). 하지만 연구소들이 이를 해결할 것이라고 추측하는 것이 합리적이며, 그렇게 되면 단순히 규모 확장 곡선을 유지하는 데 그치지 않고 모델 능력에 엄청난 향상을 가져올 수도 있다고 생각합니다.

덧붙여 말하자면, 이는 앞으로 몇 년간 연구소 간 변동성이 오늘날보다 훨씬 커질 것임을 의미하기도 합니다. 최근까지는 최첨단 기술이 공개되어 모두가 기본적으로 같은 일을 해왔습니다. (신생 연구소나 오픈 소스 프로젝트

20. 연산 효율 3배 증가의 경제적 이익은 클러스터 비용을 감안할 때 수십억 달러 이상으로 측정될 것입니다.

21. 대략 10배 정도의 향상에 해당합니다.

16. 약 GPT-3 스케일에서 3배 이상, 더 큰 스케일에서는 3배 이상.

트도 공개된 레시피 덕분에 최전선과 경쟁할 수 있었습니다.) 이제는 핵심 알고리즘 아이디어가 점점 더 독점화되고 있습니다. 연구소들의 접근법이 훨씬 더 다양해지고, 일부는 다른 곳보다 더 빠른 진전을 이룰 것으로 예상됩니다—지금은 최전선에 있는 것처럼 보이는 연구소도 데이터 장벽에 막혀 정체될 수 있고, 다른 연구소가 돌파구를 찾아 앞서 나갈 수도 있습니다. 오픈 소스는 경쟁이 훨씬 더 어려워질 것입니다. 확실히 흥미로운 상황이 될 것입니다. (그리고 만약 어느 연구소가 이를 해결한다면, 그 돌파구는 AGI의 열쇠이자 초지능의 열쇠가 될 것이며—미국이 가장 소중히 여기는 비밀 중 하나가 될 것입니다.)

제약 해소

마지막으로, 정량화하기 가장 어려우면서도 결코 덜 중요하지 않은 개선 범주가 있습니다. 제가 “제약 해소”라고 부를 부분입니다.

어려운 수학 문제를 풀라고 했을 때, 즉시 떠오르는 첫 번째 답만 내야 한다고 상상해 보세요. 가장 단순한 문제를 제외하면 어려움을 겪을 것이 분명합니다. 하지만 최근까지 우리는 LLM들이 수학 문제를 그렇게 풀게 했습니다. 대신 대부분의 사람들은 문제를 스크래치패드에서 단계별로 풀면서 훨씬 더 어려운 문제를 해결할 수 있죠. “사고의 연쇄(chain-of-thought)” 프롬프트가 LLM들에게 그 가능성을 열어주었습니다. 뛰어난 원시 능력에도 불구하고, LLM들은 명백한 제약 때문에 수학에서 훨씬 못 미치는 성과를 냈고, 작은 알고리즘 조정만으로 훨씬 더 큰 능력을 발휘할 수 있었습니다.

우리는 지난 몇 년간 모델의 “제약 해소”에서 엄청난 진전을 이루었습니다. 이는 단순히 더 나은 기본 모델을 훈련하는 것을 넘어서는 알고리즘 개선이며, 종종 사전 훈련 컴퓨팅의 일부만 사용하면서 모델 능력을 폭발적으로 끌어올립니다:

- 인간 피드백을 통한 강화 학습(RLHF). 기본 모델은 놀라운 잠재 능력을 가지고 있지만,²⁵ 다루기 매우 어렵습니다. RLHF에 대한 대중적 인식은 단지 욕설을 검열하는 것에 불과하다고 생각하지만, RLHF는 모델이 실제로 유용하고 상업적으로 가치 있게 만드는 핵심이었습니다(단순히 무작위 인터넷 텍스트를 예측하는 것이 아니라, 실제로 능력을 적용해 질문에 답하려고 시도하게 만드는 것). 이것이 ChatGPT의 마법이었고—잘 수행된 RLHF가 모델을 실제 사람들에게 처음으로 사용 가능하고 유용하게 만들었습니다. 원래 InstructGPT 논문에는 이 점이 잘 정량화되어 있는데, RLHF가 적용된 작은 모델이 인간 평가자 선호도 기준으로 RLHF가 적용되지 않은 100배 이상 큰 모델과 동등했습니다.
- 사고의 연쇄(Chain of Thought, CoT). 앞서 논의한 바와 같이, CoT는 불과 2년 전에 널리 사용되기 시작했으며, 수학 및 추론 문제에서 >10배 이상의 유효 컴퓨팅 증가 효과를 제공합니다.
- 발판(Scaffolding). CoT++라고 생각하세요: 단순히 모델에게 묻는 것만이 아니라

문제를 해결하기 위해, 한 모델이 공격 계획을 세우고, 다른 모델이 여러 가능한 해결책을 제안하며, 또 다른 모델이 그것을 비판하는 식으로 진행할 수 있다. 예를 들어, HumanEval(코딩 문제)에서는 간단한 스캐폴딩이 GPT-3.5가 스캐폴딩 없는 GPT-4보다 더 잘 수행하게 한다. SWE-Bench(실제 소프트웨어 엔지니어링 과제 해결 벤치마크)에서는 GPT-4가 약 2%만 정확히 해결할 수 있지만, Devin의 에이전트 스캐폴딩을 사용하면 14-23%로 뛰어오른다. (에이전트 해방은 아직 초기 단계이며, 이에 대해서는 나중에 더 자세히 논의할 것이다.)

22. 그리고 같은 교과서를 계속 반복해서 읽는 것은 이해가 아니라 암기만 할 가능성이 있다. 나는 많은 단어 중심 학습자(wordcels)가 수학 수업을 그렇게 통과한다고 생각한다!

- 도구: 인간이 계산기나 컴퓨터를 사용할 수 없다고 상상해보라. 우리는 이제 시작 단계에 불과하지만, ChatGPT는 이제 웹 브라우저를 사용하고, 코드를 실행하는 등의 기능을 갖추고 있다.

- 컨텍스트 길이: 모델들은 2천 토큰 컨텍스트(GPT-3)에서 3만 2천 토큰 컨텍스트(GPT-4 출시)로, 그리고 100만 토큰 이상 컨텍스트(Gemini 1.5 Pro)로 발전해왔다. 이는 매우 중요한 변화다. 예를 들어, 10만 토큰의 관련 컨텍스트를 가진 훨씬 작은 기본 모델이 4천 토큰의 관련 컨텍스트만 가진 훨씬 큰 모델보다 더 뛰어날 수 있다. 더 많은 컨텍스트는 사실상 큰 컴퓨팅 효율성 향상이다.²⁶ 더 일반적으로, 컨텍스트는 이들 모델의 많은 응용을 여는 열쇠다: 예를 들어, 많은 코딩 응용은 유용한 새 코드를 기여하기 위해 코드베이스의 큰 부분을 이해해야 하며; 또는 업무용 문서 작성에 모델을 사용할 경우, 관련된 내부 문서와 대화의 컨텍스트가 많이 필요하다. Gemini 1.5 Pro는 100만 토큰 이상의 컨텍스트를 통해 사전 자료로 사전과 문법 참고 자료를 넣는 것만으로 인터넷에 없는 저자원 언어를 처음부터 학습할 수 있었다!

- 사후 학습 개선: 현재 GPT-4는 출시 당시의 원래 GPT-4에 비해 상당히 개선되었다고 John Schulman은 말한다. 이는 잠재된 모델 능력을 해방한 사후 학습 개선 덕분이다. 추론 평가에서 상당한 향상을 이루었는데 (예: MATH에서 약 50%에서 72%로, GPQA에서 약 40%에서 50%로), LMSys 리더보드에서는 거의 100점의 옐로 점수 상승을 기록했다. 이는 Claude 3 Haiku와 훨씬 큰 Claude 3 Opus 모델 간의 옐로 차이와 비슷하며, 두 모델은 약 50배 가격 차이가 난다.

업데이트된 Gemini 1.5 백서의 그림 7을 참조하라. Gemini 1.5 Pro와 Gemini 1.5 Flash(더 저렴하고 아마도 더 작은 모델)의 혼란도 대비 컨텍스트 비교를 보여준다.

Epoch AI가 스캐폴딩(scaffolding), 도구 사용(tool use) 등과 같은 일부 기법을 조사한 결과, 이러한 기법들은 많은 벤치마크에서 일반적으로 5~30배의 유효 컴퓨팅 향상을 가져오는 것으로 나타났다. 모델을 평가하는 기관인 METR도 동일한 GPT-4 기반 모델에서 제약 해소(unhobbling)를 통해 에이전트 작업(agent tasks)에서 매우 큰 성능 향상을 발견했다. 기본 모델만 사용할 때 5%였던 성능이, 출시 후 GPT-4의 추가 학습(posttraining)으로 20%까지, 그리고 더 나은 추가 학습, 도구, 에이전트 스캐폴딩을 통해 현재 거의 40%에 이르고 있다(그림 16).

23. 흥미로운 관점 하나: 사전학습과 맥락 내 학습 사이에는 ‘중간 결손(missing-middle)’이 존재합니다. 맥락 내 학습은 놀라울 정도로 효율적이며(인간의 샘플 효율성과 경쟁), 예를 들어 Gemini 1.5 Pro 논문에서는 약 200명 미만이 사용하는 거의 인터넷에 존재하지 않는 Kalamang 언어의 교재와 사전(instructional materials)을 모델에 주었고, 모델이 영어에서 Kalamang으로 인간 수준으로 번역하는 법을 학습했습니다! 맥락 내에서는 모델이 교재로부터 인간만큼 잘 학습하지만, 단순히 그 교재를 사전학습에 집어넣는 것보다 훨씬 낫습니다. 인간은 교재 학습을 통해 단기 기억/학습을 장기 기억/장기 기술로 전환하지만, 우리는 ‘맥락 내 학습’을 ‘가중치에 복귀시키는’ 동등한 방법이 없습니다. 합성 데이터/자기 플레이/RL 등이 이를 해결하려고 합니다: 모델이 스스로 학습하고, 이를 성찰하며, 학습한 것을 연습하여 그 학습을 가중치에 증류시키는 것입니다.

24. 여기서 안드레이 카파시(Andrej Karpathy)의 강연도 참고하라.

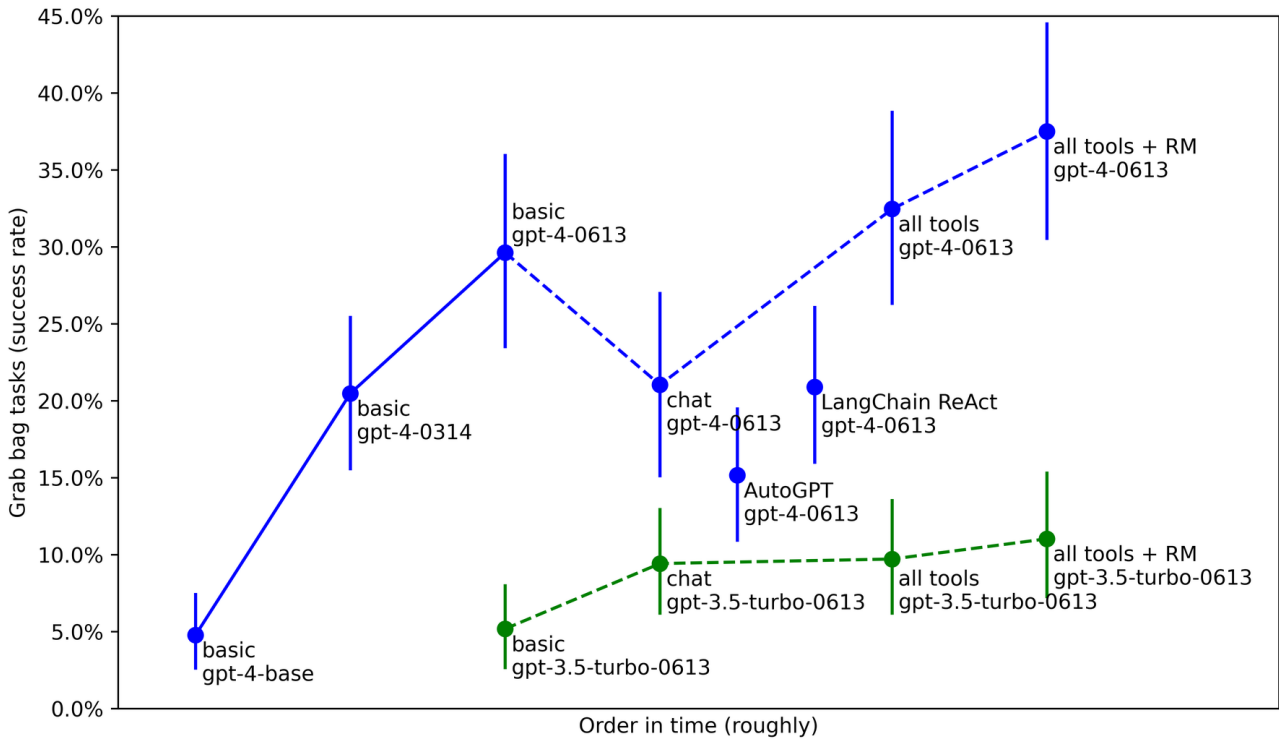


그림 16: METR의 에이전트 작업에서 더 나은 제약 해소를 통한 시간 경과별 성능. 출처: 모델 평가 및 위협 연구

이것들을 컴퓨팅 및 알고리즘 효율성과 통합된 유효 컴퓨팅 척도로 표현하기는 어렵지만, 적어도 컴퓨팅 규모 확장과 알고리즘 효율성에 근접하는 대규모 향상임은 분명하다. (이는 알고리즘 진보의 중심적 역할도 강조한다. 이미 상당한 약 0.5 규모 차수(OOM)/년의 컴퓨팅 효율성은 이야기의 일부에 불과하며, 제약 해소와 결합된 알고리즘 진보가 현재 추세에서 전체 향상의 과반수를 차지할 수도 있다.)

“제약 해소(unhobbling)”는 이러한 모델들이 실제로 유용해질 수 있게 한 매우 중요한 부분이며, 오늘날 많은 상업적 응용이 지연되는 이유도 이러한 추가적인 “제약 해소”의 필요성 때문이라고 주장할 수 있다. 실제로 오늘날 모델들은 여전히 매우 제약되어 있다! 예를 들어:

- 장기 기억이 없다.

25. 이것이 비지도 학습의 마법입니다: 다음 토큰을 더 잘 예측하고 당황도를 낮추기 위해 모델은 감정부터 복잡한 세계 모델에 이르기까지 놀랄 만큼 풍부한 내부 표현을 학습합니다. 그러나 기본 상태에서는 이 강력한 내부 표현을 단순히 인터넷 랜덤 텍스트의 다음 토큰 예측에만 사용하며, 실제 문제 해결에 최적화해 적용하지는 못합니다.

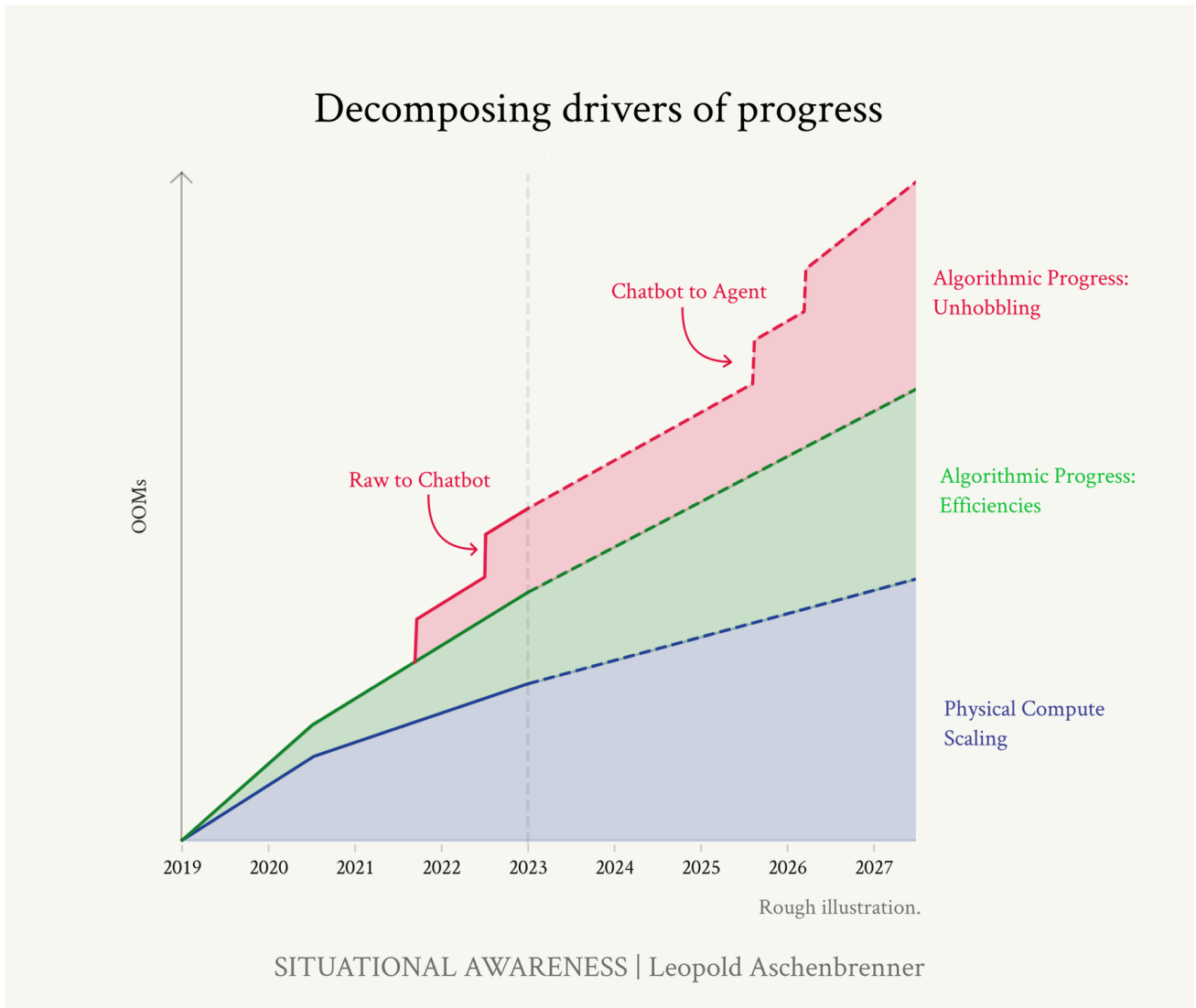


그림 17:진보 분해: 컴퓨팅, 알고리즘 효율성, 그리고 제약 해소. (대략적 설명.)

- 그들은 컴퓨터를 사용할 수 없다 (아직 매우 제한된 도구만 가지고 있다).
- 그들은 여전히 대부분 말하기 전에 생각하지 않는다. ChatGPT에게 에세이를 쓰라고 하면, 이는 인간이 처음 떠오르는 의식의 흐름으로 에세이를 쓰는 것과 같다.²⁷
- 그들은 (대부분) 짧은 문답식 대화만 할 수 있으며, 하루나 일주일 동안 문제를 고민하고, 다양한 접근법을 연구하며, 다른 사람들과 상담한 후 더 긴 보고서나 풀 리퀘스트를 작성하는 것은 불가능하다.
- 그들은 대부분 당신이나 당신의 애플리케이션에 맞춤형되어 있지 않다 (짧은 프롬프트를 가진 일반적인 챗봇 일 뿐이며, 당신의 회사와 업무에 관한 모든 관련 배경 지식을 갖추고 있지 않다).

여기서 가능성은 엄청나며, 우리는 빠르게 손쉬운 과실을 따고 있다. 이것이 중요하다: 단순히 “GPT-6 ChatGPT”만을 상상하는 것은 완전히 잘못된 것이다. 계속되는 제약 해소와 함께

26. 업데이트된 Gemini 1.5 백서의 그림 7 참고: Gemini 1.5 Pro와 Gemini 1.5 Flash(더 저렴하고 아마도 더 작은 모델)의 당향도와 문맥 길이 비교.

27. 사람들이 이 문제를 해결하기 위해 노력하고 있다.

진전이 이루어지면, 개선은 GPT-6 + RLHF에 비해 획기적인 변화가 될 것입니다. 2027년쯤에는 챗봇이 아니라, 동료 같은 에이전트에 더 가까운 무언가를 갖게 될 것입니다.

챗봇에서 에이전트-동료로

앞으로 몇 년간 야심 찬 제약 해소는 어떻게 보일까요? 제가 생각하는 핵심 요소는 세 가지입니다:

1. “온보딩 문제” 해결

GPT-4는 많은 사람들의 업무 중 상당 부분을 수행할 수 있는 기본적인 지능을 갖추고 있지만, 마치 5분 전에 막 들어온 똑똑한 신입사원과 같습니다. 관련된 맥락이 없고, 회사 문서나 슬랙 기록을 읽거나 팀원들과 대화를 나누거나 회사 내부 코드베이스를 이해하는 데 시간을 들이지 않았습니다. 똑똑한 신입사원도 도착한 지 5분 만에 는 별로 쓸모가 없지만, 한 달이 지나면 꽤 유용해집니다! 예를 들어 매우 긴 컨텍스트를 통해, 새로운 인간 동료를 온보딩하듯 모델을 “온보딩”하는 것이 가능할 것 같습니다. 이것만으로도 엄청난 돌파구가 될 것입니다.

1. 테스트 시점 컴퓨팅 과잉(추론/오류 수정/장기 문제 해결을 위한 시스템 II)

현재 모델은 기본적으로 짧은 작업만 수행할 수 있습니다: 질문을 하면 답을 줍니다. 하지만 이것은 매우 제한적입니다. 인간이 하는 대부분의 유용한 인지 작업은 장기적입니다—5분 만에 끝나는 것이 아니라, 몇 시간, 며칠, 몇 주, 또는 몇 달이 걸립니다.

5분만 어려운 문제를 생각할 수 있는 과학자는 어떤 과학적 돌파구도 만들 수 없습니다. 요청받았을 때 단일 함수의 뼈대 코드만 작성할 수 있는 소프트웨어 엔지니어도 별로 쓸모가 없습니다—소프트웨어 엔지니어는 더 큰 과제를 받고, 계획을 세우고, 관련 코드베이스 부분을 이해한 후에 작업을 진행합니다.

기술 도구를 사용하여 다양한 모듈을 작성하고 점진적으로 테스트하며, 오류를 디버깅하고 가능한 해결책의 공간을 탐색한 후, 결국 몇 주간의 작업이 집약된 대규모 풀 리퀘스트를 제출하는 식이다. 그리고 계속된다.

본질적으로, 테스트 시점의 컴퓨팅에 큰 과잉 부담이 있다. 각 GPT-4 토큰을 문제를 생각할 때의 내적 독백의 한 단어로 생각해 보라. 각 GPT-4 토큰은 꽤 똑똑하지만, 현재는 일관성 있게 사고의 연쇄를 위해 약 수백 개 토큰 정도만 효과적으로 사용할 수 있다(사실상 문제나 프로젝트에 대해 몇 분간의 내적 독백/사고만 할 수 있는 것과 같다).

만약 수백만 개의 토큰을 사용해 정말 어려운 문제나 더 큰 프로젝트에 대해 사고하고 작업할 수 있다면 어떨까?

토큰 수 내가 어떤 일에 투자하는 시간과 동등한... 수백 개 몇 분 ChatGPT (현재 수준) 수천 개 30분 +1 규모 차수(OOM) 테스트 시점 컴퓨팅 만 개 반나절 +2 규모 차수 십만 개 일주일 +3 규모 차수 수백만 개 수개월 +4 규모 차수

표 2: 사람이 분당 약 100 토큰을 생각하고 주당 40시간 일한다고 가정할 때, 토큰 수로 “모델이 사고하는 시간”을 인간 시간으로 환산한 것.

“토큰당” 지능이 동일하다고 해도, 똑똑한 사람이 몇 분을 투자하는 것과 몇 달을 투자하는 것의 차이일 것이다. 나는 당신은 모르겠지만, 몇 분보다 몇 달 동안 할 수 있는 일이 훨씬, 훨씬, 훨씬 많다. 만약 모델이 “몇 분에 해당하는 사고가 아니라 몇 달에 해당하는 사고와 작업을 할 수 있게” 된다면, 엄청난 능력 향상이 열릴 것이다. 여기에는 수많은 규모 차수(OOM)의 과잉 부담이 존재한다.

현재 모델들은 아직 이것을 할 수 없다. 최근 장기 문맥 관련 진전이 있었지만, 이 긴 문맥은 주로 토큰 소비에만 효과적이며, 토큰 생성에는 그렇지 않다—어느 정도 시간이 지나면 모델이 엉뚱한 방향으로 가거나 멈춰 버린다. 아직 일정 시간 동안 떨어져서 작업할 수 있는 능력은 없다.

문제나 프로젝트 자체만으로는²⁸

본질적으로, 우리는 모델에게 어려운 장기 프로젝트를 추론할 수 있게 하는 일종의 시스템 II 외부 루프²⁹를 가르쳐야 한다.

만약 이 외부 루프를 가르치는 데 성공한다면, 몇 단락짜리 짧은 챗봇 답변 대신, 모델이 문제를 생각하고 도구를 사용하며 다양한 접근법을 시도하고 연구하며 작업을 수정하고 다른 사람과 협력하여 스스로 큰 프로젝트를 완성하는 수백만 단어의 흐름(읽는 것보다 더 빠르게 들어오는)을 상상해 보라.

다른 머신러닝 분야에서 테스트 시점과 학습 시점 컴퓨팅의 균형 조정. 보드 게임용 AI 시스템과 같은 다른 분야에서는, 더 많은 테스트 시점 컴퓨팅(추론 시점 컴퓨팅이라고도 함)을 사용하여 학습 컴퓨팅을 대체할 수 있음이 입증되었다.

28. 이해가 된다—왜 더 긴 시간 범위의 추론과 오류 수정 기술을 배웠겠는가? 하지만 테스트 시점 컴퓨팅을 해제하는 것은 상대적으로 작은 “제약 해소” 알고리즘 효율성의 승리일 수 있다. 아마도 소량의 강화학습(RL)이 모델이 오류를 수정하는 법을 배우도록 도울 수 있다 (“음, 그건 맞지 않는 것 같은데, 다시 확인해 보자”), 계획을 세우고 가능한 해결책을 탐색하는 등. 어떤 면에서 모델은 이미 대부분의 원시 능력을 가지고 있으며, 단지 이를 모두 결합하기 위해 몇 가지 추가 기술을 배워야 할 뿐이다.

29. 시스템 I과 시스템 II는 현재 LLM의 능력—특히 그 한계와 어리석은 실수—및 RL과 제약 해제의 가능성을 고려하는 데 유용한 사고 방식입니다. 익숙한 비유로, 운전할 때 대부분은 자동조종(시스템 I, 현재 모델들이 주로 하는 것) 상태이나, 복잡한 공사 구간이나 낯선 교차로를 만나면 잠시 대화를 멈추고 곰곰이 생각하는(시스템 II) 행동을 합니다. 만약 시스템 I만 사용해야 한다면(현 모델에 가깝게), 큰 어려움이 있을 것입니다. 시스템 II 추론 루프를 만드는 것이 핵심 해제점입니다.

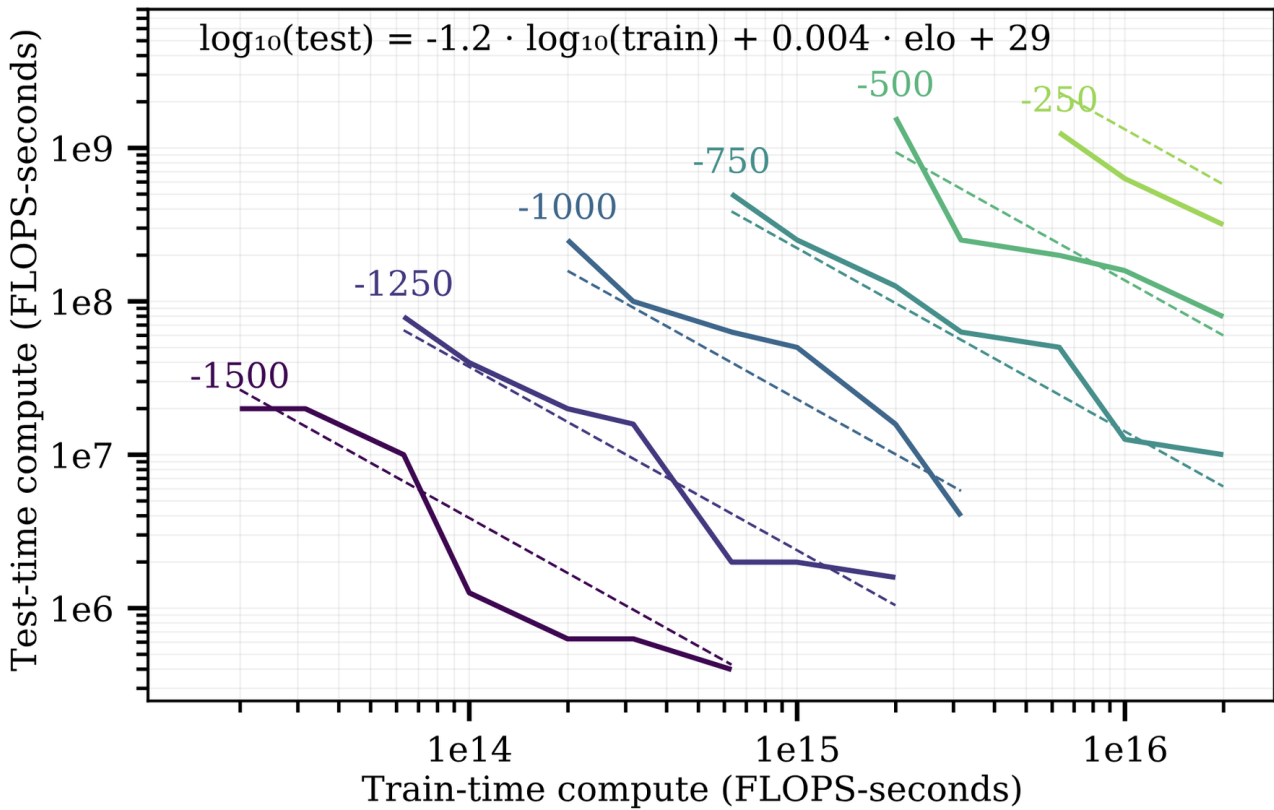


Fig. 9. The trade-off between train-time compute and test-time compute. Each dotted line gives the minimum train-test compute required for a certain Elo on a 9×9 board

그림 18: Jones (2021): 더 작은 모델도 더 많은 테스트 시점 컴퓨팅(“더 많은 생각 시간”)을 제공하면 Hex 게임에서 훨씬 더 큰 모델만큼 잘할 수 있다. 이 분야에서는 테스트 시점에 약 1.2 규모 차수(OOM) 더 많은 컴퓨팅을 사용하여 학습 컴퓨팅이 약 1 OOM 더 많은 모델과 동등한 성능을 얻을 수 있음을 발견했다.

시스템 I 대 시스템 II는 현재 대형 언어 모델(LLM)의 능력—제한과 어리석은 실수—과 강화학습 및 제약 해소로 가능해질 수 있는 것을 생각하는 유용한 방법이다. 이렇게 생각해 보라: 운전할 때 대부분은 자동 조종(시스템 I, 현재 모델들이 주로 하는 것) 상태이다. 하지만 복잡한 공사 구역이나 새로운 교차로를 만나면, 동승자에게 대화를 잠시 멈추라고 부탁하고 실제로 무슨 일이 일어나고 무엇을 해야 할지 생각한다. 만약 시스템 I만으로 생활해야 한다면(오늘날 모델에 더 가까움), 많은 어려움을 겪을 것이다. 시스템 II 추론 루프를 만드는 능력이 핵심 해제이다.

(각주 내용 없음)

만약 우리 경우에도 비슷한 관계가 성립한다면, 테스트 시점 컴퓨팅을 +4 규모 차수(OOM)만큼 해제할 수 있다면, 이는 사전 학습 컴퓨팅을 +3 OOM만큼 해제하는 것과 동등할 수 있습니다. 즉, 대략 GPT-3와 GPT-4 사이의 도약과 비슷한 수준입니다. (이 “제약 해소” 문제를 해결하는 것은 거대한 규모 차수 확장과 같습니다.)

1. 컴퓨터 사용하기

이것은 세 가지 중 아마도 가장 직관적인 부분입니다. 현재 ChatGPT는 기본적으로 텍스트로 소통할 수 있는 고립된 상자 안에 앉아 있는 인간과 같습니다. 초기 제약 해소 개선은 모델이 개별 고립된 도구를 사용하는 법을

가르치지만, 멀티모달 모델이 등장하면 곧 한 번에 이 모든 것을 할 수 있을 것으로 예상합니다. 즉, 모델이 인간처럼 컴퓨터를 사용할 수 있게 될 것입니다.

이는 Zoom 통화에 참여하고, 온라인에서 조사하며, 사람들과 메시지를 주고받고 이메일을 보내고, 공유 문서를 읽고, 앱과 개발 도구를 사용하는 것을 의미합니다. (물론, 모델이 장기적인 작업 루프에서 이를 최대한 활용하려면 테스트 시점 컴퓨팅 해제가 함께 이루어져야 합니다.)

이 과정이 끝나면, 우리는 마치 바로 투입 가능한 원격 근로자와 매우 흡사한 무언가를 얻게 될 것입니다. 회사에 합류하여 신입 직원처럼 온보딩되고, Slack에서 당신과 동료들에게 메시지를 보내며 소프트웨어를 사용하고, 풀 리퀘스트를 만들고, 큰 프로젝트가 주어지면 인간이 몇 주간 독립적으로 프로젝트를 완수하는 것과 동등한 모델의 역할을 수행하는 에이전트입니다. 이를 실현하려면 GPT-4보다 다소 나은 기본 모델이 필요할 수 있지만, 그리 크게 뛰어나지 않아도 됩니다—모델이 아직도 명백하고 기본적인 방식으로 제약받는 부분을 고치는데 많은 잠재력이 있습니다.

(이것이 어떻게 보일지에 대한 아주 초기의 예는 Devin입니다. Devin은 완전 자동화된 소프트웨어 엔지니어를 만드는 과정에서 모델의 “에이전시 잔여/테스트 시점 컴퓨팅 잔여”를 해제하는 초기 프로토타입입니다. Devin이 실제로 얼마나 잘 작동하는지는 모르겠고, 이 데모는 아직 매우)

적절한 챗봇 → 에이전트 제약 해소가 가져올 결과에 비해 제한적이지만, 곧 다가올 그런 종류의 것을 미리 보여주는 유용한 예고편이다.)

그런데 제약 해소의 중심성은 상업적 응용 측면에서 다소 흥미로운 “음속 돌파” 효과를 낼 수 있다. 지금과 드롭인 원격 근로자 사이의 중간 모델들은 워크플로우를 변경하고 경제적 가치를 통합하고 도출하기 위한 인프라를 구축하는 데 엄청난 수고가 필요할 것이다. 드롭인 원격 근로자는 훨씬 더 쉽게 통합될 것이다—그냥, 원격으로 수행할 수 있는 모든 작업을 자동화하기 위해 그들을 투입하기만 하면 된다. 수고가 제약 해소보다 더 오래 걸릴 가능성이 있어 보인다. 즉, 드롭인 원격 근로자가 많은 작업을 자동화할 수 있을 때쯤에는 중간 모델들이 아직 완전히 활용되고 통합되지 않았을 것이므로, 생성되는 경제적 가치의 도약은 다소 불연속적일 수 있다.

다음 4년

숫자를 종합해 보면, 우리는 GPT-4 이후 4년 동안 대략 GPT-2에서 GPT-4로의 도약과 비슷한 규모의 도약을 2027년 말까지 기대할 수 있다.

- GPT-2에서 GPT-4로는 대략 4.5-6 규모 차수(OOM)의 기본 유효 컴퓨팅 스케일업(물리적 컴퓨팅과 알고리즘 효율성), 그리고 주요한 “제약 해소” 이득(기본 모델에서 챗봇으로)이 있었다.

- 이후 4년 동안은 3-6 OOM의 기본 유효 컴퓨팅 스케일업(물리적 컴퓨팅과 알고리즘 효율성)—최선 추정치는 약 5 OOM—과 “제약 해소”에 의해 잠금 해제되는 유용성과 응용 분야의 단계적 변화(챗봇에서 에이전트/드롭인 원격 근로자로)가 기대된다. 30 위에 설명된 물리적 컴퓨팅과 알고리즘 효율성 스케일업에 관한 최선 추정 과정과 단순화된 병렬 처리 고려를 적용하면 이것을 관점에서 보면, GPT-4 훈련에 3개월이 걸렸다고 가정할 때,

2027년에는 선도 AI 연구소가 GPT-4 수준 모델을 1분 만에 훈련할 수 있을 것이다.³⁰ 유효 컴퓨팅 규모 차수 스케일업은 (실제로는 “하루에 1440개(6024) GPT-4 수준 모델” 또는 이와 유사한 모습일 수 있다).

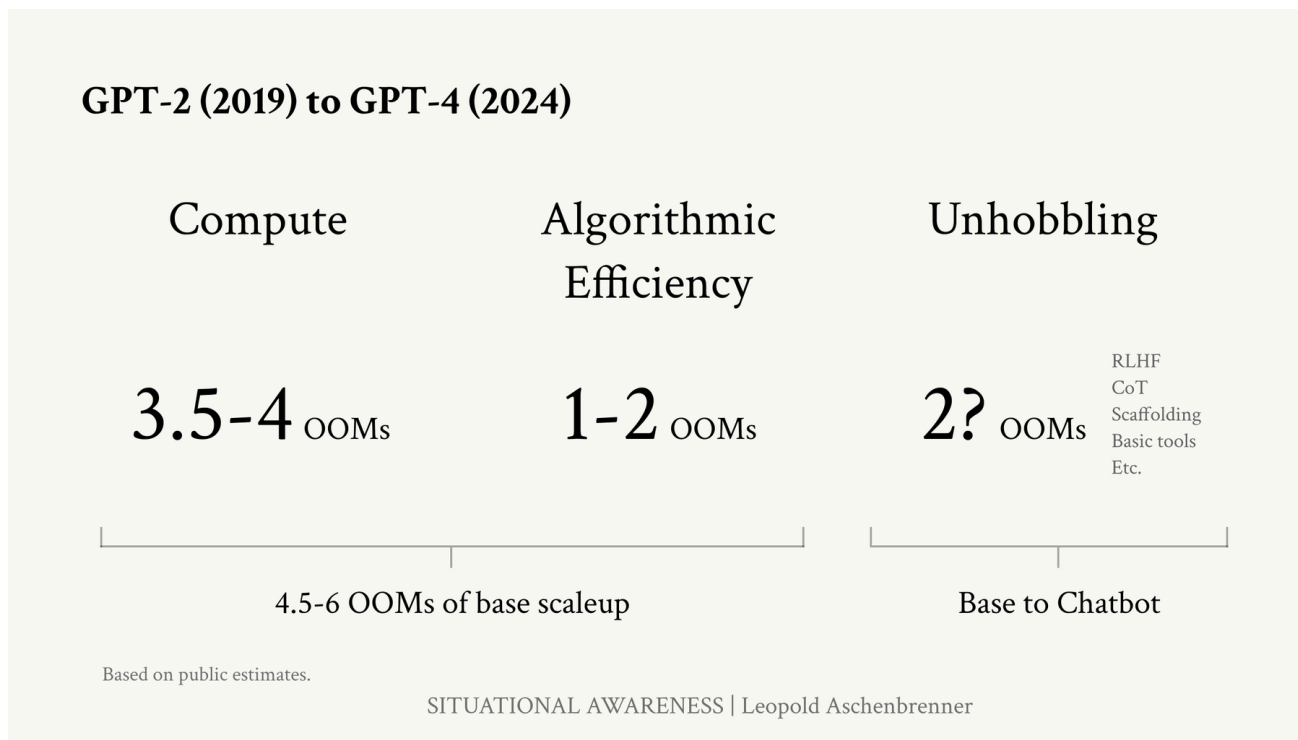


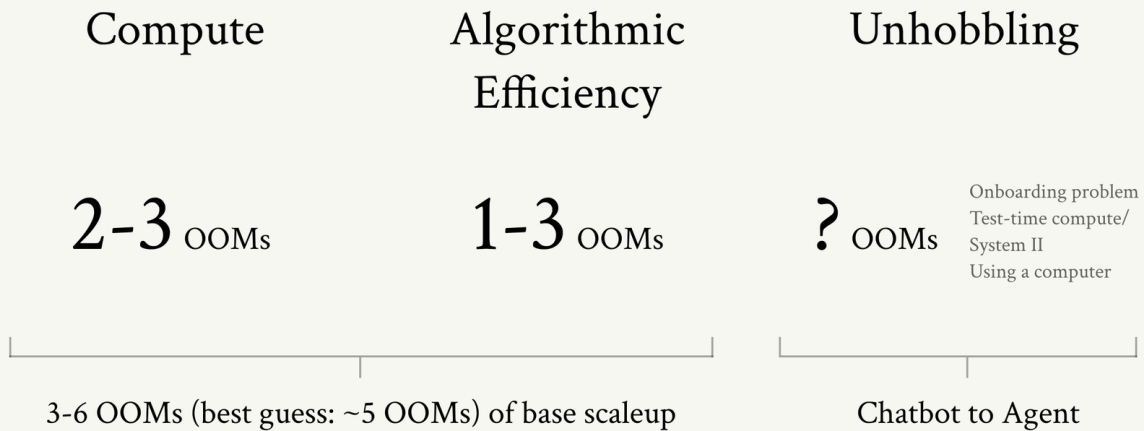
그림 19:GPT-4 이전 4년간 진보의 동인에 대한 추정치 요약과 GPT-4 이후 4년간 우리가 기대해야 할 사항.

극적으로 변할 것이다.

그것이 우리를 어디로 데려갈까?

30. 위에서 묘사한 물리 컴퓨트 및 알고리즘 효율 향상에 관한 최선 추정 가정과 병렬성 고려를 단순화하면(실제로 “하루에 1440(6024)대 GPT-4급 모델” 형태일 수 있음).

2023-2027 (Projection)



Based on public estimates.

SITUATIONAL AWARENESS | Leopold Aschenbrenner

그림 20: 규모 차수(OOM) 계산 요약. GPT-2에서 GPT-4로의 발전은 우리를 약 유치원생 수준에서 약 똑똑한 고등학생 수준으로 데려갔다; 몇 개의 일관된 문장을 간신히 출력하던 수준에서 고등학교 시험을 만점으로 통과하고 유용한 코딩 조수 역할을 할 수 있는 수준으로. 이는 엄청난 도약이었다. 만약 우리가 다시 한 번 이 정도의 지능 격차를 극복한다면, 그것이 우리를 어디로 데려갈까?³¹ 우리는 매우, 매우 멀리 갈 것이라고 놀라지 말아야 한다. 아마도 그것은 박사 학위 소지자와 해당 분야 최고의 전문가들을 능가하는 모델로 우리를 이끌 것이다.

(이것을 생각하는 한 가지 멋진 방법은 현재 AI 발전 추세가 대략 아이 발달 속도의 3배 정도라는 것이다. 당신의 3배 속도 아이는 이제 막 고등학교를 졸업했으며, 곧 당신의 일자리를 차지할 것이다!)

다시 말하지만, 매우 중요한 점은 단순히 믿을 수 없을 만큼 똑똑한 ChatGPT를 상상하지 말라는 것이다: 제약 해소(unhobbling)로 인한 이득은 이것이 오히려 원격 근무자로서, 추론하고 계획하며 오류를 수정할 수 있고 당신과 당신 회사에 대해 모든 것을 알고 문제를 독립적으로 해결할 수 있는 믿을 수 없을 만큼 똑똑한 에이전트 처럼 보일 것임을 의미한다.

몇 주 동안 독립적으로 진행되고 있다.

우리는 2027년까지 범용 인공지능(AGI)에 도달할 것으로 예상된다. 이 AI 시스템들은 기본적으로 모든 인지 작업을 자동화할 수 있을 것이다(원격으로 수행할 수 있는 모든 작업을 생각해 보라).

명확히 하자면—오차 범위는 크다. 데이터가 부족해지거나, 데이터 장벽을 돌파하는 데 필요한 알고리즘 혁신이 예상보다 어려울 경우 진전이 멈출 수 있다. 제약 해소가 기대만큼 진행되지 않아 단순한 전문가용 챗봇 수준에 머물 수도 있다. 아마도 10년간의 추세선이 깨지거나, 딥러닝 확장이 이번에는 실제로 한계에 부딪힐 수도 있다. (또는 알고리즘 혁신, 심지어 테스트 시점의 컴퓨팅 과잉을 해방시키는 단순한 제약 해소가 패러다임 전환이 되어 상황을 가속화하고 AGI를 더 일찍 도래하게 할 수도 있다.)

31. 물론 현재 우리가 가진 벤치마크는 모두 포화될 것입니다. 그러나 이는 어려울 만큼 어려운 벤치마크를 만드는 것이 어려움의 반영일 뿐입니다.

어쨌든 우리는 규모 차수(OOM)를 빠르게 통과하고 있으며, 비전문적인 믿음 없이도 단순히 직선 추세를 외삽하는 것만으로 2027년까지 진정한 AGI 가능성을 매우 심각하게 받아들여야 한다.

요즘 많은 이들이 AGI를 단순히 매우 훌륭한 챗봇 정도로 하향 정의하는 게임에 참여하는 것 같다. 내가 말하는 AGI란 나나 내 친구들의 일을 완전히 자동화할 수 있고, AI 연구자나 엔지니어의 업무를 완전히 수행할 수 있는 AI 시스템을 의미한다. 아마도 로보틱스 같은 일부 분야는 기본적으로 더 오래 걸릴 수 있다. 그리고 의료나 법률 같은 사회적 도입은 사회적 선택이나 규제로 인해 쉽게 지연될 수 있다. 그러나 모델이 AI 연구 자체를 자동화할 수 있게 되면, 그것만으로도 충분하다—강력한 피드백 루프를 촉발하기에 충분하며—우리는 매우 빠르게 추가 진전을 이룰 수 있다. 자동화된 AI 엔지니어들이 남은 모든 병목 현상을 해결하며 모든 것을 완전히 자동화할 것이다. 특히 수백만 명의 자동화된 연구자들이 10년간의 추가 알고리즘 진보를 1년 이하로 압축할 가능성이 매우 높다. AGI는 곧 뒤따를 초지능의 작은 맛에 불과할 것이다. (다음 장에서 더 다룬다.)

어쨌든, 급격한 진전 속도가 줄어들 것이라고 기대하지 마라. 추세선은 무해해 보이지만, 그 함의는...

강렬하다. 이전 세대들처럼, 새로운 세대의 모델들은 대부분의 관찰자들을 놀라게 할 것이다; 곧 박사 학위 소지자들이 며칠씩 걸릴 매우 어려운 과학 문제들을 모델들이 해결할 때, 모델들이 당신의 컴퓨터에서 당신의 일을 척척 해낼 때, 수백만 줄의 코드를 처음부터 작성할 때, 그리고 매년 혹은 2년마다 이 모델들이 창출하는 경제적 가치가 10배씩 증가할 때, 사람들은 믿기 어려워할 것이다. 공상과학은 잊어라, 규모 차수(OOM)를 세어라: 이것이 우리가 기대해야 할 것이다. 범용 인공지능(AGI)은 더 이상 먼 미래의 환상이 아니다. 단순한 딥러닝 기법을 확장하는 것이 성공했으며, 모델들은 배우고자 할 뿐이고, 우리는 2027년 말까지 또 다른 100,000배 이상의 성장을 이루려 하고 있다. 그들이 우리보다 더 똑똑해지는 것도 머지않았다.

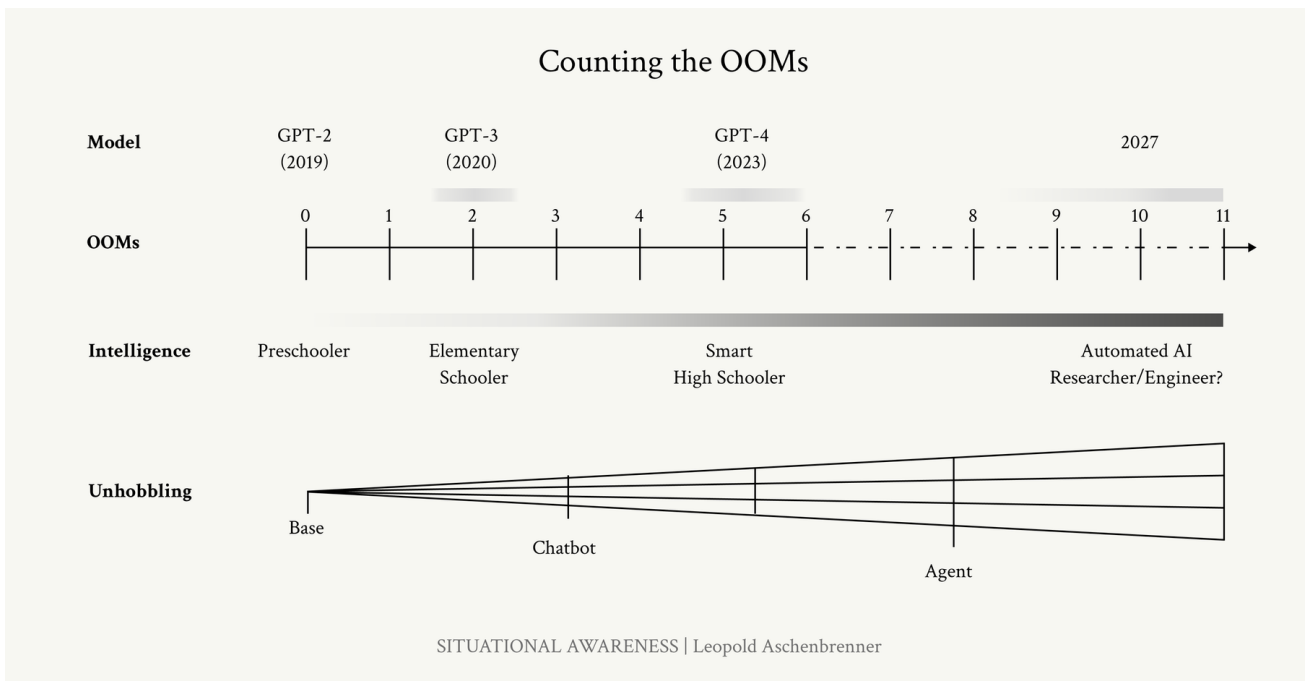


그림 21: GPT-4는 시작에 불과하다— 4년 후 우리는 어디에 있을까? 딥러닝 진보의 빠른 속도를 과소평가하는 실수를 하지 말라(이는 GAN의 진보로 잘 나타난다).

부록. 규모 차수(OOM)를 가로지르는 경쟁: 이번 10년 아니면 실패

나는 예전에 범용 인공지능(AGI)에 대한 짧은 시간표에 대해 더 회의적이었다. 한 가지 이유는 이번 10년에 너무 많은 AGI 확률 질량을 집중하는 것이 비합리적으로 보였기 때문이다(“우리가 특별하다”는 전형적인 오류처럼 보였다). 나는 AGI를 얻기 위해 무엇이 필요한지에 대해 불확실해야 하며, 이는 AGI를 얻을 시기에 대해 훨씬 더 “분산된” 확률 분포를 가져야 한다고 생각했다.

그러나 내 생각이 바뀌었다: 결정적으로, AGI를 얻기 위해 무엇이 필요한지에 대한 우리의 불확실성은 연도가 아니라 유효 컴퓨팅의 규모 차수(OOM)에 관한 것이어야 한다.

우리는 이번 10년 동안 규모 차수를 가로지르는 경쟁을 하고 있다. 무어의 법칙이 전성기였던 시절에도 1~1.5 OOM/10년 수준이었다. 나는 앞으로 4년 내에 약 5 OOM, 그리고 이번 10년 전체로는 10 OOM 이상을 달성할 것으로 추정한다.



그림 22: 유효 컴퓨팅 규모 확장에 대한 대략적인 예상. 이번 10년 동안 규모 차수를 가로지르는 경쟁을 해왔으며, 2030년대 초반 이후에는 느린 진전이 예상된다.

본질적으로, 우리는 이번 10년 동안 일회성 이득을 거두는 거대한 규모 확장의 한가운데에 있으며, 이후 규모 차수(OOM)를 통한 진전은 수배 느려질 것이다.⁴⁵ 만약 이 규모 확장이 향후 5~10년 내에 AGI에 도달하지 못한다면, 그것은 아마도...

아직 갈 길이 멀다.

- **지출 규모 확대:** 모델에 백만 달러를 쓰는 것은 한때 터무니없었지만, 이 10년 말까지는 1,000억 달러 또는 1조 달러 규모의 클러스터를 갖게 될 가능성이 높다. 그 이상으로 가는 것은 어려울 것이다; 이미 이것이 사실상 가능한 한계(대기업이 감당할 수 있는 수준과 GDP 대비 비율 모두에서)이다. 이후에는 연간 2% 정도의 빙하 같은 실질 GDP 성장률만이 이를 증가시킬 수 있다.
- **하드웨어 발전:** AI 하드웨어는 무어의 법칙보다 훨씬 빠르게 발전해왔다. 이는 AI 작업 부하에 특화된 칩을 개발해왔기 때문이다. 예를 들어, CPU에서 GPU로 전환했고, 트랜스포머에 맞게 칩을 조정했으며, 전통적인 슈퍼컴퓨팅의 fp⁶⁴/fp32에서 H100의 fp8과 같이 훨씬 낮은 정밀도 숫자 형식으로 내려갔다. 이는 큰 발전이지만,

45. 유사하게, 반대상 경제성장 이론에서 과학 투자를 GDP 대비 1%에서 20%로 늘려도 성장률이 영원히 높아지진 않지만, “수준 효과(level effect)”가 크게 작용해 수십 년간 성장 가속화에 기여합니다.

10년 말까지는 완전히 특화된 AI 전용 칩이 나올 것이며, 무어의 법칙을 넘어서는 추가적인 발전은 많지 않을 것이다.

• 알고리즘 진보: 다가오는 10년 동안 AI 연구소들은 수십억 달러를 알고리즘 연구개발에 투자할 것이며, 전 세계에서 가장 똑똑한 사람들이 이 분야에 몰릴 것이다; 작은 효율성 개선부터 새로운 패러다임까지, 많은 저위험 과실을 따먹을 것이다. 우리는 어떤 확고한 한계에 도달하지는 않을 것이지만(“제약 해소”는 유한할 가능성이 크다), 최소한 개선 속도는 느려질 것이다. 이는 급격한 성장(금전 및 인적 자본 투자)이 필연적으로 둔화되기 때문이다(예: 대부분의 우수한 STEM 인재가 이미 AI에 몰려 있을 것이다). (그럼에도 불구하고, 이것이 가장 예측하기 어려운 부분이며, 위 그래프에서 2030년대 규모 차수(OOM)에 대한 불확실성의 주요 원천이다.)

이 모든 것을 종합하면, 우리는 앞으로 10년 동안 이후 수십 년보다 훨씬 더 많은 규모 차수를 빠르게 통과하고 있다. 어쩌면 이것으로 충분해서 곧 범용 인공지능(AGI)을 얻을 수도 있고, 아니면 길고 느린 고난의 행군이 될 수도 있다. 당신과 나는 AGI까지의 중간 시간에 대해 합리적으로 의견이 다를 수 있다.

AGI 달성이 얼마나 어려울지에 따라 다르겠지만— 지금 우리가 규모 차수를 빠르게 통과하고 있는 점을 감안하면, 당신이 생각하는 일반적인 AGI 도달 시기는 아마도 이번 10년대 후반쯤일 것이다.

64. 태양광과 배터리도 완전히 불가능한 대안은 아니지만 천연가스보다는 다소 더 어렵습니다. Casey Handmer가 계산한 지구를 태양광 패널로 덮는 방안은 흥미롭습니다: “현재 GPU로 글로벌 태양광 데이터센터는 약 1,500억 명 인간과 동등한 연산력을 가지며, 컴퓨터가 인간 뇌 효율에 근접하면 5경에 달하는 AI 영혼들을 지원할 수 있습니다.”

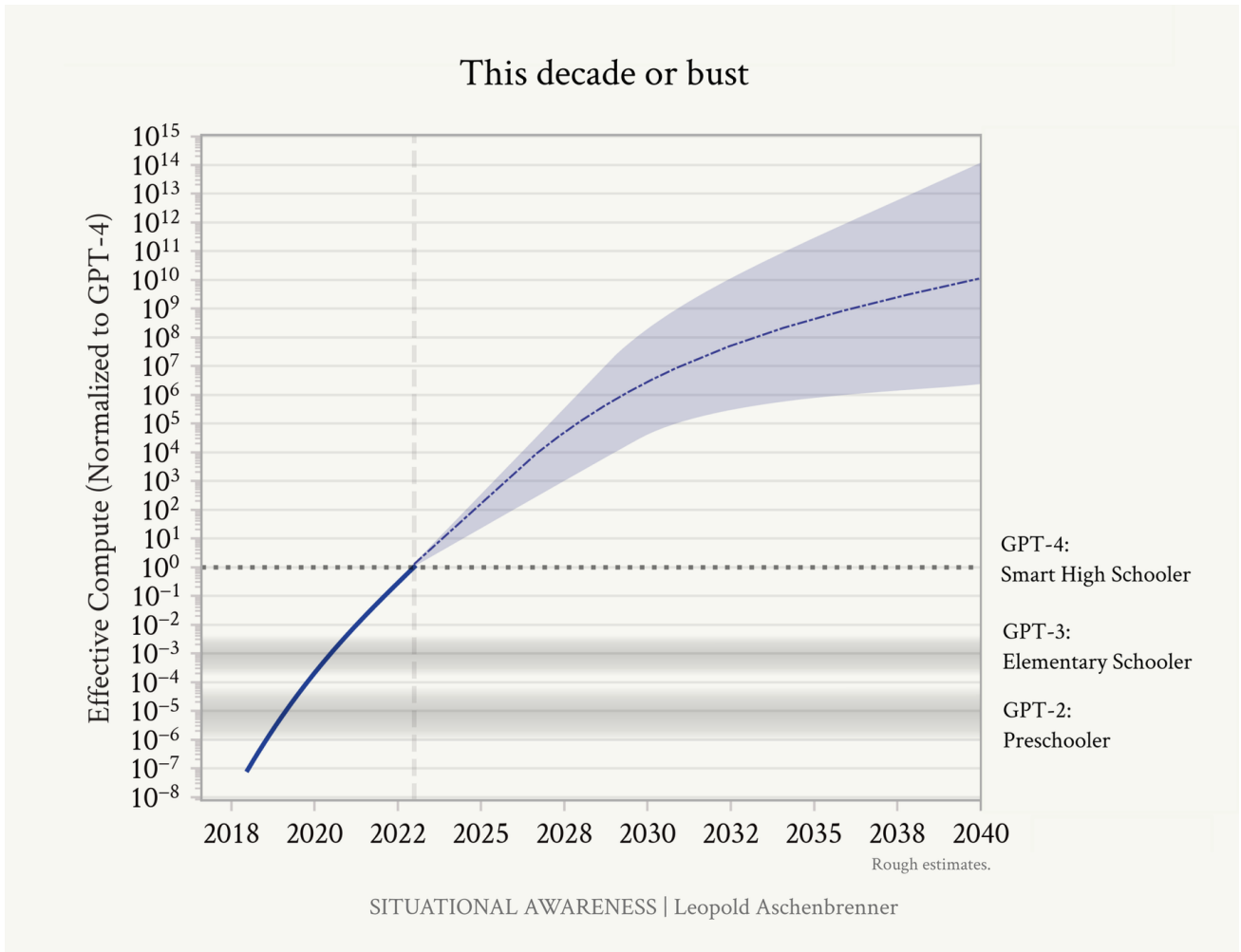


그림 23:매튜 바넷은 컴퓨팅과 생물학적 한계만을 고려한 이와 관련된 멋진 시각화를 제공한다.

II. 범용 인공지능(AGI)에서 초지능으로: 지능 폭발

AI 발전은 인간 수준에서 멈추지 않을 것이다. 수억 대의 AGI가 AI 연구를 자동화하여 10년에 걸친 알고리즘 효율성의 진보(5+ 규모 차수)를 1년으로 압축할 수 있다. 우리는 인간 수준에서 급격히 초인간적 AI 시스템으로 도약할 것이다. 초지능의 힘과 위험은 극적일 것이다.

초지능 기계란 아무리 영리한 인간의 모든 지적 활동을 훨씬 능가할 수 있는 기계로 정의할 수 있다. 기계 설계가 이러한 지적 활동 중 하나이므로, 초지능 기계는 더 나은 기계를 설계할 수 있다; 그러면 의심할 여지 없이 ‘지능 폭발’이 일어나 인간의 지능은 훨씬 뒤처지게 될 것이다. 따라서 최초의 초지능 기계는 인간이 만들어야 할 마지막 발명품이다.

i. j. 굿 (1965)

폭탄과 초무기 일반적인 상상에서 냉전의 공포는 주로 로스앨러모스와 원자폭탄 발명에서 비롯된 것으로 여겨진다. 그러나 폭탄 자체만으로는 과대평가된 측면이 있다. 원자폭탄에서 초무기인 수소폭탄으로의 발전이 아마도 똑같이 중요했다.

도쿄 공습에서는 수백 대의 폭격기가 수천 톤의 재래식 폭탄을 도시에 투하했다. 그해 후반 히로시마에 투하된 리틀 보이싱 단일 장치에서 비슷한 파괴력을 발휘했다. 그러나 불과 7년 후, 텔러의 수소폭탄은 그 위력을 다시

천 배로 증폭시켰다—제2차 세계대전 전체에서 투하된 모든 폭탄의 폭발력을 합친 것보다 더 강력한 단일 폭탄이었다.

폭탄은 더 효율적인 폭격 캠페인이었다. 수퍼는 국가를 전멸시키는 장치였다.³²

범용 인공지능(AGI)과 초지능도 마찬가지일 것이다.

AI 발전은 인간 수준에서 멈추지 않을 것이다. 처음에는 최고의 인간 게임에서 학습한 후, AlphaGo는 스스로 대국을 시작했고—빠르게 초인적인 수준에 도달하여 인간이 결코 생각해내지 못할 매우 창의적이고 복잡한 수를 두었다.

우리는 이전 글에서 AGI로 가는 경로를 논의했다. AGI를 얻으면, 한 번 더—혹은 두세 번 더—크랭크를 돌려 AI 시스템이 초인적 수준, 즉 훨씬 초인적인 수준이 될 것이다. 그들은 당신이나 나보다 질적으로 훨씬 더 똑똑해질 것이며, 아마도 당신이나 내가 초등학생보다 질적으로 더 똑똑한 것과 비슷할 것이다.

초지능으로의 도약은 현재의 빠르지만 연속적인 AI 발전 속도에서도 충분히 놀라울 것이다(만약 GPT-4에서 4년 내에 AGI로 도약할 수 있다면, 그 후 4년 또는 8년이 더 지나면 무엇이 올지 상상해 보라). 하지만 AGI가 AI 연구 자체를 자동화한다면 훨씬 더 빠를 수도 있다.

AGI를 얻으면 단 하나의 AGI만 있는 것이 아니다. 나중에 수치를 자세히 설명하겠지만, 그때쯤이면 추론용 GPU 클러스터 덕분에 수백만 대(아마도 1억 인간 등가체, 그리고 곧 10배 이상의 인간 속도도 가능할 것이다)를 운영할 수 있을 것이다. 그들이 아직 사무실을 돌아다니거나 커피를 만들 수는 없더라도, 컴퓨터에서 기계학습 연구를 수행할 수 있을 것이다. 선도 AI 연구소의 수백 명 연구원과 엔지니어 대신, 10만 배가 넘는 인원이 밤낮없이 알고리즘 혁신에 열중할 것이다. 네, 재귀적 자기 개선이지만, 공상과학은 필요 없다; 그들은 단지 현재의 알고리즘 효율성 추세(현재 약 연간 0.5 규모 차수(OOM))를 가속화하기만 하면 된다.

자동화된 AI 연구는 아마도 인간 10년 분량의 알고리즘 진보를 1년도 채 안 되는 기간에 압축할 수 있을 것이다(이것도 보수적인 추정이다). 이는 5 이상 규모 차수(OOM)이며, GPT-2에서—

32. 그리고 냉전의 많은 왜곡들(예: 다니엘 엘스버그의 책)은 단순히 A-폭탄을 H-폭탄으로 교체하면서 핵 정책과 전쟁 계획을 대규모 능력 증가에 맞게 조정하지 않은 데서 비롯되었다.



Matthew Barnett
@MatthewJBar

...

My own basic calculations suggest that, given the potential for increased investment and hardware progress, we could very soon move through a large fraction of the remaining compute gap between the current frontier models and the literal amount of computation used by evolution.



7:54 PM · Mar 26, 2024 · 3,968 Views

그림 24: 자동화된 AI 연구는 알고리즘 진보를 가속화하여 GPT-4 크기 급증과 더불어 AGI 위에—유치원생에서 똑똑한 고등학생으로의 질적 도약과 같은—이미 전문가 AI 연구원/엔지니어만큼 똑똑한 AI 시스템 위에 추가되는 도약을 이룰 수 있다. 1년 내에 5단계 이상의 유효 컴퓨팅(OOM) 향상을 가져올 수 있다. 지능 폭발이 끝날 무렵 우리가 갖게 될 AI 시스템은 인간보다 훨씬 더 똑똑할 것이다.

여러 가지 그럴듯한 병목 현상이 있다—실험을 위한 제한된 컴퓨팅, 인간과의 상호 보완성, 알고리즘 진보가 점점 어려워지는 점 등—이를 다루겠지만, 어느 것도 확실히 속도를 늦출 만큼 충분해 보이지 않는다.

우리가 깨닫기도 전에, 우리는 초지능을 손에 넣게 될 것이다—인간보다 훨씬 더 똑똑한 AI 시스템으로, 우리가 이해조차 시작할 수 없는 새롭고 창의적이며 복잡한 행동을 할 수 있다—어쩌면 수십억 개의 작은 문명일 수도 있다. 그들의 힘도 막대할 것이다. 초지능을 적용하면—

지능을 다른 분야의 연구개발에 투입한다면, 폭발적인 진보는 단지 머신러닝 연구에 국한되지 않고 곧 로봇틱스를 해결하며 수년 내에 과학과 기술의 다른 분야에서 극적인 도약을 이루고 산업적 폭발이 뒤따를 것이다. 초지능은 결정적인 군사적 우위를 제공할 가능성이 높으며, 헤아릴 수 없는 파괴력을 펼칠 것이다. 우리는 인류 역사상 가장 강렬하고 불안정한 순간 중 하나에 직면하게 될 것이다.

AI 연구 자동화

우리는 모든 것을 자동화할 필요는 없다—단지 AI 연구만 자동화하면 된다. AGI의 변혁적 영향에 대한 흔한 반론은 AI가 모든 것을 수행하기 어렵다는 것이다. 예를 들어 로보틱스를 보라, 회의론자들은 말한다; AI가 박사 수준의 인지능력을 갖추더라도 그것은 까다로운 문제일 것이다. 또는 많은 물리적 실험실 작업과 인간 실험이 필요한 생물학 연구개발 자동화를 예로 들 수 있다.

하지만 AI가 AI 연구를 자동화하는 데 로보틱스나 많은 것들이 필요하지는 않다. 선도 연구소의 AI 연구원과 엔지니어의 업무는 완전히 가상으로 수행할 수 있으며, 실제 세계의 병목 현상에 부딪히지 않는다(물론 컴퓨팅 자원에 의해 제한되겠지만, 이는 나중에 다룰 것이다). 그리고 AI 연구원의 업무는 큰 틀에서 보면 꽤 직관적이다: 머신러닝 문헌을 읽고 새로운 질문이나 아이디어를 제시하며, 그 아이디어를 시험할 실험을 구현하고, 결과를 해석하며 이를 반복한다. 이 모든 것은 현재 AI 능력의 단순한 외삽으로도 2027년 말까지 최고 인간 수준에 도달하거나 그 이상에 쉽게 이를 수 있는 영역이다.³³

지난 10년간 가장 큰 머신러닝 돌파구들이 얼마나 직관적이고 간단한 해킹이었는지 강조할 가치가 있다: “아, 그냥 정규화(normalization)를 좀 추가해라”(LayerNorm/BatchNorm) 또는 “ $f(x)$ 대신 $f(x)+x$ 를 해라”(잔차 연결) 또는 “구현 버그를 고쳐라”(Kaplan → Chinchilla 스케일링 법칙). AI 연구는 자동화할 수 있다. 그리고 AI 연구 자동화만으로도 비범한 피드백 루프를 촉발할 수 있다.³⁴

우리는 수백만 개의 자동화된 AI 연구자 복제본을 실행할 수 있을 것이며(곧 인간 속도의 10배 이상으로도 가능해질 것이다). 2027년까지는 GPU 함대가 수천만 대에 이를 것으로 예상된다. 훈련 클러스터만 해도 약 3 규모 차수(OOM) 이상 커질 것이며, 이미 1,000만 개 이상의 A100급 GPU에 해당하는 규모가 될 것이다. 추론(인퍼런스) 함대는 이보다 훨씬 더 클 것이다. (이와 관련된 내용은 IIIa. 조 단위 달러 클러스터 경쟁에서 더 다룬다.)

이렇게 되면 자동화된 AI 연구자 복제본을 수백만 개, 아마도 1억 명의 인간 연구자에 해당하는 규모로 밤낮없이 실행할 수 있다. 여기에는 인간이 분당 100 토큰을 ‘생각한다’는 가정(대략적인 규모 추정치로, 예를 들어 당시의 내적 독백을 생각해 보라)과 역사적 추세 및 최첨단 모델의 토큰당 추론 비용에 대한 Chinchilla 스케일링 법칙을 외삽하는 가정이 포함된다.³⁵ 또한 일부 GPU는 실험과 새로운 모델 훈련에 예약해 두어야 한다. 전체 계산은 각주에 있다.³⁶

33. AI 연구원의 업무는 AI 연구소의 AI 연구원들이 정말 잘 알고 있는 일이기도 하다—그래서 그들에게는 모델을 그 업무에 잘 맞도록 최적화하는 것이 특히 직관적일 것이다. 그리고 그들의 연구와 연구소의 경쟁력을 가속화하기 위해 그렇게 할 엄청난 인센티브가 있을 것이다.

34. 덧붙여 말하자면, 이는 AI로 인한 위험의 순서와 관련해 중요한 점을 시사한다. 사람들이 흔히 지적하는 AI 위험 모델은 AI 시스템이 새로운 생물무기를 개발하여 치명적인 위험을 초래하는 것이다. 그러나 AI 연구가 생물학 연구개발보다 자동화하기 더 직관적이라면, 극단적인 AI 생물 위험이 발생하기 전에 지능 폭발이 먼저 일어날 수 있다. 이는 예를 들어 AI가 통제 불능 상태에 빠지기 전에 “생물학적 경고 신호”를 기대할 수 있는지 여부와 관련해 중요하다.

35. 앞서 언급했듯이, GPT-4 API 비용은 GPT-3 출시 당시보다 오늘날 더 저렴하다—이는 추론 효율성 향상이 모델이 훨씬 더 강력해져도 추론 비용을 대체로 일정하게 유지할 만큼 빠르게 진행되고 있음을 시사한다. 마찬가지로, GPT-4 출시 이후 1년 만에 추론 비용에서 큰 개선이 있었다; 예를 들어, 현재 버전인 Gemini 1.5 Pro는 원래 GPT-4보다 성능이 뛰어나면서도 비용은 약 10배 저렴하다. 또한 Chinchilla 스케일링 법칙을 고려하면 모델 크기, 즉 추론 비용은 훈련 비용의 제곱근에 비례하여 증가한다. 즉, 유효 컴퓨팅 규모 확장의 절반 정도의 규모 차수(OOM)만큼 증가한다. 그러나 이전 글에서는 알고리즘 효율성이 컴퓨팅 규모 확장과 거의 같은 속도로 발전하여 유효 컴퓨팅 규모 확장의 절반 정도를 차지한다고 제안했다. 만약 이러한 알고리즘 개선이 추론 효율성으로도 이어진다면, 알고리즘 효율성은 추론 비용의 단순 증가를 상쇄할 것이다. 실제로 훈련 컴퓨팅 효율성은 종종, 그러나 항상은 아니지만, 추론 효율성 향상으로 이어진다. 게다가 훈련 효율성과 별개로 추론 효율성에서의 많은 개선도 존재한다. 따라서 대략적인 규모 차원에서 최첨단 모델의 토큰당 비용이 대체로 비슷하게 유지된다고 가정하는 것은 무리가 아니다. (물론 더 많은 토큰, 즉 더 많은 테스트 시점 컴퓨팅을 사용할 것이다. 하지만 이는 인간 등가치를 분당 100 토큰으로 가격 책정하는 계산에 이미 포함되어 있다.)

다른 관점에서 보면, 2027년의 추론 함대를 고려할 때 매일 인터넷 전체 분량의 토큰을 생성할 수 있을 것이다.
³⁷ 어쨌든 정확한 숫자는 단순한 타당성 입증 이상의 의미는 크지 않다.

더욱이, 우리의 자동화된 AI 연구자들은 곧 인간 속도보다 훨씬 빠르게 작동할 수 있을 것입니다:

- 일부 추론 페널티를 감수함으로써, 더 적은 수의 복사본을 실행하는 대신 더 빠른 직렬 속도로 실행하는 거리를 할 수 있습니다. (예를 들어, 자동화된 연구자 100만 개 복사본을 “단지” 실행함으로써 약 5배 인간 속도에서 약 100배 인간 속도로 갈 수 있습니다.³⁸) Jacob Steinhardt는 k^3 개의 병렬 복사본을 k^2 배 더 빠른 단일 모델로 대체할 수 있다고 추정합니다. 이는 타일링 방식의 추론 트레이드오프 수학에 기반하며(이론적으로 k 가 100 이상일 때도 작동). 초기 속도가 이미 약 5배 인간 속도(예: GPT-4 출시 당시 속도)라고 가정하면, 이 추론 페널티를 감수함으로써(k =약 5) 약 100만 개의 자동화된 AI 연구자를 약 100배 인간 속도로 실행할 수 있습니다.

- 더 중요한 것은, 자동화된 AI 연구자들이 처음으로 작업하는 알고리즘 혁신은 10배 또는 100배 속도 향상입니다. Gemini 1.5 Flash는 원래 출시된 GPT-4보다 약 10배 빠르며,³⁹ 단지 1년 만에 비슷한 추론 벤치마크 성능을 제공합니다. 만약 수백 명의 인간 연구자가 1년 동안 찾을 수 있는 알고리즘 속도 향상이 이 정도라면, 자동화된 AI 연구자들은 매우 빠르게 유사한 성과를 낼 수 있을 것입니다. 이 출처는 Flash의 처리량이 약 6배 GPT-4 Turbo이며, GPT-4 Turbo는 원래 GPT-4보다 빠르다고 벤치마크합니다. 지연 시간도 대략 10배 빠를 가능성이 높습니다.

즉, AI 연구 자동화가 가능해지기 시작한 직후, 각각 100배 인간 속도로 작업하는 1억 명의 자동화된 연구자가 있을 것으로 예상할 수 있습니다. 이들은 각각 몇 일 만에 1년치 작업을 수행할 수 있습니다. 오늘날 선도적인 AI 연구소에서 1배 인간 속도로 작업하는 수백 명의 미미한 인간 연구자와 비교할 때, 연구 노력의 증가는 엄청날 것입니다.

이것은 기존 알고리즘 진보 추세를 극적으로 가속화하여 10년간의 발전을 1년으로 압축할 수 있습니다. 자동화된 AI 연구가 AI 진보를 강력히 가속화하기 위해 전혀 새로운 것을 가정할 필요는 없습니다. 앞서 살펴본 수치를 통해, 알고리즘 진보가 지난 10년간 딥러닝 발전의 핵심 동력이었음을 알 수 있었습니다; 알고리즘 효율성만으로 연간 약 0.5 규모 차수(OOM)의 추세선이 있었고, 제약 해소로 인한 추가적인 큰 알고리즘 이득도 있었습니다. (저는 알고리즘 진보의 중요성이 많은 사람들에게 과소평가되었다고 생각하며, 이를 제대로 인식하는 것이 지능 폭발 가능성을 이해하는 데 중요하다고 봅니다.)

우리의 수백만 자동화된 AI 연구자들이(곧 10배 또는 100배 인간 속도로 작업하며) 인간 연구자들이 10년간 찾을 알고리즘 진보를 1년으로 압축할 수 있을까요?

대신에? 그럼 1년에 5단계 이상의 규모 차수(OOM)가 될 것이다.

여기서 1억 명의 주니어 소프트웨어 엔지니어 인턴을 상상하지 마라(우리는 그들을 더 일찍, 앞으로 몇 년 안에 확보할 것이다!). 진짜 자동화된 AI 연구자들은 매우 똑똑할 것이며—그들의 순수한 양적 우위 외에도, 자동화된 AI 연구자들은 인간 연구자들에 비해 다른 엄청난 이점들을 가질 것이다:

36. GPT4T는 1,000 토큰당 약 0.03달러이다. 우리는 1,000만 대 이상의 A100급 GPU를 가질 것으로 가정했으며, A100급 GPU 1대당 시간당 비용은 약 1달러이다. API 비용을 토큰 생성량으로 환산하면, 1,000만 대 GPU * 1달러/시간 * 33,000 토큰/달러 = 시간당 약 1조 토큰 생성이 가능하다. 인간은 분당 100 토큰을 생각하므로 시간당 6,000 토큰이다. 1조 토큰/시간을 6,000 토큰/인간-시간으로 나누면 약 2억 명의 인간 등가치가 된다—즉, 2억 명의 인간 연구자를 밤낮없이 실행하는 것과 같다. (실험용 컴퓨팅을 위해 GPU 절반을 예약해도 1억 명 인간 연구자 등가치를 얻는다.)

- 그들은 지금까지 작성된 모든 머신러닝 논문을 읽을 수 있을 것이고, 연구소에서 수행된 모든 이전 실험을 깊이 있게 생각할 수 있으며, 각자의 복제본으로부터 병렬로 학습하여 수천 년에 해당하는 경험을 빠르게 축적할 수 있을 것이다. 그들은 어떤 인간보다도 훨씬 깊은 머신러닝 직관을 개발할 수 있을 것이다.
- 그들은 수백만 줄의 복잡한 코드를 쉽게 작성할 수 있고, 전체 코드베이스를 맥락 속에 유지하며, 수십 년(또는 그 이상)에 해당하는 인간 시간이 걸리는 코드의 버그와 최적화를 반복해서 점검할 수 있을 것이다. 그들은 작업의 모든 부분에서 매우 뛰어난 역량을 발휘할 것이다.
- 각 자동화된 AI 연구자를 개별적으로 훈련시킬 필요가 없을 것이다(실제로 1억 명의 새로운 인간 직원을 훈련하고 적응시키는 것은 어려울 것이다). 대신, 한 명만 가르치고 적응시키면 복제할 수 있다. (정치적 문제, 문화 적응 등을 걱정할 필요도 없으며, 그들은 밤낮으로 최고의 에너지와 집중력으로 일할 것이다.)
- 엄청난 수의 자동화된 AI 연구자들이 맥락을 공유할 수 있을 것이며(아마도 서로의 잠재 공간에 접근하는 것도 가능할 것이다), 이는 인간 연구자들에 비해 훨씬 더 효율적인 협업과 조정을 가능하게 할 것이다.
- 그리고 물론, 초기 자동화된 AI 연구자들이 아무리 똑똑하더라도, 우리는 곧 더 높은 규모 차수(OOM) 도약을 이뤄내어, 자동화된 AI 연구에 훨씬 더 능숙한 더 똑똑한 모델들을 만들어낼 수 있을 것이다.

자동화된 알렉 라드포드(Alec Radford)를 상상해 보라—1억 명의 자동화된...

자동화된 Alec Radford들.⁴⁰ OpenAI의 거의 모든 연구자가 10명의 Alec Radford가 있다면, 100명이나 1,000명, 심지어 1백만 명이 인간 속도의 10배 또는 100배로 작동한다면, 그들은 매우 빠르게 많은 문제를 해결할 수 있을 것이라고 동의할 것입니다. 여러 다른 병목 현상(잠시 후 더 설명)에도 불구하고, 알고리즘 진보의 10년치를 1년 만에 압축하는 것은 매우 그럴듯해 보입니다. (1백만 배 더 많은 연구 노력을 통한 10배 가속은 오히려 보수적인 추정입니다.)

그것만으로도 5+ 규모 차수(OOM)가 됩니다. 5 OOM의 알고리즘 성과는 GPT-2에서 GPT-4로의 도약과 비슷한 규모 확장으로, 능력 면에서 약 유치원생에서 똑똑한 고등학생으로의 도약과 같습니다. 그런 질적 도약이 범용 인공지능(AGI) 위에, Alec Radford 위에 더해진다고 상상해 보십시오.

우리가 AGI에서 초지능으로 매우 빠르게, 어쩌면 1년 내에 도달할 가능성은 매우 그럴듯합니다.

가능한 병목 현상

이 기본 이야기는 놀라울 정도로 강력하며 철저한 경제 모델링 작업에 의해 뒷받침되지만, 자동화된 AI 연구 지능 폭발을 아마도 늦출 몇 가지 실제적이고 그럴듯한 병목 현상들이 있습니다.

여기서 요약은 드리고, 관심 있는 분들을 위해 아래 선택적 섹션에서 더 자세히 논의하겠습니다:

- 제한된 컴퓨팅: AI 연구는 단지 좋은 아이디어, 사고, 수학만 필요한 것이 아니라, 아이디어에 대한 경험적 신호를 얻기 위한 실험 실행도 필요합니다. 자동화된 연구 노동을 통한 1백만 배 더 많은 연구 노력은 1백만 배 더 빠른 진보를 의미하지 않습니다. 왜냐하면 컴퓨팅은 여전히 제한적일 것이고, 실험을 위한 제한된 컴퓨팅이 병목 현상이 될 것이기 때문입니다. 그럼에도 불구하고, 1,000,000배 가속은 아닐지라도, 자동화된 AI 연구자들이 컴

37. 앞 각주에서 시간당 약 1조 토큰, 즉 하루 24조 토큰을 추정했다. 이전 글에서는 공개 중복 제거된 CommonCrawl 데이터가 약 30조 토큰임을 언급했다.

퓨팅을 적어도 10배 더 효과적으로 사용할 수 있다고 상상하기 어렵지 않습니다. 그들은 엄청난 ML 직관을 얻을 수 있을 것입니다(전체 ML 문헌과 이전에 실행된 모든 실험을 내면화했기 때문입니다) 그리고

정확히 어떤 실험을 실행할지 알아내고, 최적으로 구성하며, 최대한의 정보 가치를 얻기 위해 수 세기 상당의 사고 시간을 쏟을 수 있다; 작은 실험조차도 실행 전에 수 세기 상당의 엔지니어 시간을 투자해 버그를 피하고 첫 시도에 제대로 수행할 수 있다; 가장 큰 성과에 집중해 컴퓨팅을 절약하는 트레이드오프를 할 수 있다; 그리고 수많은 소규모 실험을 시도할 수 있다(그때쯤이면 유효 컴퓨팅 규모가 커져 “소규모”가 1년에 10만 개의 GPT-4 수준 모델을 훈련시켜 아키텍처 혁신을 시도할 수 있는 의미가 된다). 일부 인간 연구자와 엔지니어는 같은 컴퓨팅을 써도 다른 이들보다 10배 더 많은 진전을 이룰 수 있는데—이것은 자동화된 AI 연구자에게는 더욱 적용될 것이다. 나는 이것이 가장 중요한 병목이라고 생각하며, 아래에서 더 깊이 다룬다.

- 상호보완성/긴 꼬리: 경제학의 고전적 교훈(바움올의 성장병 참조)은 만약 어떤 작업의 70%를 자동화할 수 있다면 어느 정도 이득은 있지만, 나머지 30%가 곧 병목이 된다는 것이다. 완전 자동화에 미치지 못하는 어떤 것—예를 들어 정말 좋은 코파일럿—에 대해서는 인간 AI 연구자가 여전히 주요 병목으로 남아 알고리즘 진보 속도의 전반적 증가가 상대적으로 작을 것이다. 게다가 AI 연구 자동화에 필요한 능력에는 긴 꼬리가 있을 가능성이 높다—AI 연구자의 작업 중 마지막 10%는 특히 자동화하기 어려울 수 있다. 이것은 지능 폭발의 가속을 다소 완화할 수 있지만, 내 최선의 추측은 이것이 단지 몇 년 정도 지연시킬 뿐이라는 것이다. 아마도 2026/27년 모델 속도가 원시 자동화 연구자의 수준이고, 완전 자동화에 이르기 위해서는 1~2년 더 제약 해소, 좀 더 나은 모델, 추론 속도 향상, 문제 해결이 필요하며, 결국 2028년쯤 10배 가속(그리고 10년 말까지 초지능)을 달성할 것이다.

- 알고리즘 진보의 본질적 한계: 알고리즘 효율성에서 추가로 5 규모 차수(OOM)를 근본적으로 달성할 수 없을 까? 나는 그렇지 않다고 본다. 분명 상한선은 있겠지만,⁴¹ 지난 10년간 5 OOM을 달성했다면 앞으로도 최소한 10년 정도의 진보는 기대할 수 있을 것이다.

가능할 것이다. 보다 직접적으로, 현재의 아키텍처와 학습 알고리즘은 여전히 매우 원시적이며, 훨씬 더 효율적인 방식이 가능할 것으로 보인다. 생물학적 참조 클래스도 훨씬 더 효율적인 알고리즘이 가능함을 뒷받침한다.

- 아이디어를 찾기 점점 어려워져서, 자동화된 AI 연구자들은 현재의 진보 속도를 가속화하기보다는 단지 유지할 뿐이다: 한 가지 반론은 자동화된 연구가 유효 연구 노력을 크게 증가시키겠지만, 아이디어를 찾는 것이 점점 더 어려워진다는 것이다. 즉, 오늘날 연구소에서 수백 명의 최고 연구자가 0.5 규모 차수(OOM)/년의 진보를 유지하는 데 충분하지만, 저위험 과실이 고갈됨에 따라 그 진보를 유지하려면 점점 더 많은 노력이 필요해질 것이고, 따라서 1억 명의 자동화된 연구자들은 단지 진보를 유지하는 데 필요한 수준일 뿐이라는 것이다. 나는 이 기본 모델이 맞다고 생각하지만, 경험적 데이터는 맞지 않는다: 연구 노력의 증가 규모—백만 배—는 진보를 유지하는 데 필요한 연구 노력 증가의 역사적 추세보다 훨씬, 훨씬 크다. 경제 모델링 용어로는, 자동화로 인한 연구 노력 증가가 진보를 일정하게 유지하는 데 딱 맞을 것이라는 가정은 기묘한 ‘칼날 가장자리 가정’이다.

- 아이디어를 찾기 점점 어려워지고 수확 체감이 발생하여 지능 폭발이 빠르게 사그라든다: 위의 반론과 관련하여, 자동화된 AI 연구자들이 초기 진보의 폭발을 이끌더라도, 빠른 진보가 지속될 수 있을지는 알고리즘 진보에

38. Jacob Steinhardt는 모델의 k^3 배 병렬 복제를 k^2 배 더 빠른 단일 모델로 대체할 수 있다고 추정했으며, 이는 타일링 방식의 추론 트레이드오프 수학에 기반(이론상 $k=100$ 이상에도 적용 가능)합니다. 초기 속도가 이미 약 인간 속도의 5배(GPT-4 출시 시점 속도 기준)라면, 이 추론 페널티(k 약 5)를 감안해도 약 100배 빠른 속도로 약 100만 대 자동화 AI 연구자를 운영할 수 있습니다.

대한 수확 체감 곡선의 형태에 달려 있다. 다시 말해, 경험적 증거를 가장 잘 해석하면 지수는 폭발적/가속적 진보에 유리하게 나타난다. 어쨌든, 100명에서 수억 명의 AI 연구자로의 일회성 대규모 증가는 적어도 여러 규모 차수의 알고리즘 진보에 대해 수확 체감을 극복할 가능성이 크며, 물론 무한히 지속될 수는 없다.

전반적으로, 이러한 요인들은 다소 속도를 늦출 수 있다: 가장 극단적인 지능 폭발 버전(예: 하룻밤 사이)은 그럴 듯하지 않다. 그리고 이들은 다소 더 긴...

준비 단계(아마도 진정한 자동화된 Alec Radford들이 등장하기 전, 더 느린 초기 자동화 연구자들로부터 1~2년 정도 더 기다려야 할지도 모른다). 하지만 이들이 매우 빠른 지능 폭발을 배제하지는 않는다. 완전 자동화된 AI 연구자에서부터 엄청나게 초인적인 AI 시스템으로 전환되는 데 1년—최대 몇 년, 어쩌면 몇 달밖에 걸리지 않을 것이라는 것이 우리의 주된 기대다.

아래 다양한 병목 현상에 대한 심층 논의를 건너뛰고 싶다면, 여기 클릭하여 다음 섹션으로 넘어가라.

실험용 제한된 컴퓨팅(선택 사항, 심층 내용)

알고리즘 진보의 생산 함수는 두 가지 상호 보완적인 생산 요소를 포함한다: 연구 노력과 실험용 컴퓨팅. 수백만 명의 자동화된 AI 연구자들은 인간 AI 연구자들보다 더 많은 컴퓨팅 자원을 실험에 사용할 수 없을 것이다; 아마도 그들은 작업이 끝나기만을 기다리고 있을지도 모른다.

이것이 아마도 지능 폭발에 가장 중요한 병목 현상일 것이다. 궁극적으로 이것은 정량적인 문제다—얼마나 큰 병목 현상인가? 전반적으로 나는 1억 명의 Alec Radford들이 실험용 컴퓨팅의 한계 생산성을 최소 10배는 증가시킬 수 없다는 것을 믿기 어렵다(따라서 진보 속도도 10배 가속될 것이다):

- 적은 컴퓨팅 자원으로도 할 수 있는 일이 많다. 대부분의 AI 연구는 소규모로 실험해보고, 그 후 스케일링 법칙을 통해 외삽하는 방식으로 진행된다. (많은 주요 역사적 돌파구들은 매우 적은 컴퓨팅 자원만 필요했다. 예를 들어, 원래의 Transformer는 단지 8개의 GPU로 며칠간 훈련되었다.) 그리고 앞으로 4년 내에 약 5 규모 차수(OOM)의 기본 스케일업이 이루어질 것을 감안하면, “소규모”는 GPT-4 규모를 의미할 것이다—자동화된 AI 연구자들은 연간 훈련 클러스터에서 10만 건의 GPT-4 수준 실험과 수천만 건의 GPT-3 수준 실험을 실행할 수 있을 것이다. (그것은 매우 많은 양이다)

잠재적 돌파구가 될 새로운 아키텍처들을 테스트할 수 있을 것입니다!

- 많은 컴퓨팅 자원은 최종 사전학습 실행의 대규모 검증에 투입됩니다—연간 주요 제품의 한계 효율성 향상에 대해 충분히 높은 신뢰도를 확보하는 것이죠—하지만 지능 폭발에서 규모 차수를 빠르게 넘나들고 있다면, 경제적으로 운영하며 정말 큰 성과에만 집중할 수 있습니다.

- 이전 글에서 논의했듯이, 비교적 적은 컴퓨팅으로 모델의 제약을 해소하는 데서도 엄청난 이득을 얻을 수 있습니다. 이런 경우에는 대규모 사전학습 실행이 필요하지 않습니다. 지능 폭발은 자동화된 AI 연구가 시작되면서 출발할 가능성이 매우 높습니다. 예를 들어, 제약 해소를 통해 몇 차례 규모 차수(OOM)의 성과를 내는 강화학습(RL) 방식을 발견하는 식으로 말이죠(그 후 본격적인 경쟁이 시작됩니다).

39. 해당 출처는 Flash의 처리량이 GPT-4 Turbo 대비 약 6배라고 벤치마크하며, GPT-4 Turbo는 원래 GPT-4보다 빠릅니다. 지연 시간 또한 대략 10배 빠른 가능성이 높습니다.

40. Alec Radford는 OpenAI에서 수많은 중요한 진보를 이끈 매우 재능 있고 다작하는 연구자/엔지니어이지만 다소 저평가되고 있습니다.

- 자동화된 AI 연구자들이 효율성을 발견하면 더 많은 실험을 수행할 수 있게 됩니다. 앞서 논의한 바와 같이, 동등한 MATH 성능을 내면서 2년 만에 거의 1000배 저렴해진 추론 비용과, 지난 1년간 10배 향상된 일반 추론 효율성은 단순한 인간 수준의 알고리즘 진보에서 비롯된 것입니다. 자동화된 AI 연구자들이 가장 먼저 할 일은 이와 유사한 성과를 빠르게 찾아내는 것이며, 이는 다시 새로운 강화학습 접근법에 대해 100배 더 많은 실험을 가능하게 할 것입니다. 또는 관련 분야에서 유사한 성능을 내는 더 작은 모델을 빠르게 만들 수도 있습니다(이전 글에서 논의한 GPT-4 대비 거의 100배 저렴한 Gemini Flash 참고). 이는 다시 이 작은 모델들로 훨씬 더 많은 실험을 수행할 수 있게 해줍니다(예를 들어 다양한 강화학습 방식을 시도하는 데 활용할 수 있겠죠). 아마도 다른 잠재력도 있을 것입니다. 예를 들어, 자동화된 AI 연구자들은 모든 추론 GPU를 활용할 수 있는 훨씬 더 나은 분산 학습 방식을 빠르게 개발할 수 있을지도 모릅니다(거기서만도 아마 10배 이상의 컴퓨팅 향상이 있을 것입니다). 더 일반적으로, 그들이 발견하는 훈련 효율성 향상 한 차수(OOM)마다 한 차수의...

실험을 수행할 수 있는 더 많은 유효 컴퓨팅.

- 자동화된 AI 연구자들은 훨씬 더 효율적일 수 있다. 첫 시도에 정확히 맞추기만 하면 얼마나 많은 실험을 줄일 수 있는지 과소평가하기 어렵다—복잡한 버그 없이, 정확히 무엇을 실행할지 더 신중하게 선택하는 등. 1000명의 자동화된 AI 연구자가 한 달에 해당하는 시간 동안 코드를 점검하고 정확한 실험을 맞추기 위해 노력하는 것을 상상해 보라. AI 연구소 동료들에게 이 점을 물어봤는데, 그들도 동의했다: 불필요한 버그를 피하고, 첫 시도에 맞추며, 정보 가치가 높은 실험만 실행할 수 있다면 대부분 프로젝트에서 컴퓨팅을 3배에서 10배까지 쉽게 절약할 수 있을 것이다.

- 자동화된 AI 연구자들은 훨씬 더 나은 직관을 가질 수 있다.

- 최근에 최첨단 연구소의 인턴과 대화했는데; 그들은 지난 몇 달 동안 주로 많은 실험을 제안했지만, 그들의 지도교수(선임 연구원)는 이미 결과를 예측할 수 있어서 실험할 필요가 없다고 말하는 경험이 지배적이었다고 했다. 선임 연구원의 수년간 무작위 실험과 모델 조작 경험이 어떤 아이디어가 통할지—또는 통하지 않을지에 대한 직관을 갈고닦았다. 마찬가지로, 우리의 AI 시스템도 ML 실험에 대해 초인적인 직관을 쉽게 가질 수 있을 것 같다—그들은 전체 머신러닝 문헌을 읽었고, 모든 다른 실험 결과에서 학습하며 깊이 사고할 수 있고, 수백만 건의 ML 실험 결과를 예측하도록 훈련받을 수 있다. 그리고 아마도 그들이 처음 하는 일 중 하나는 "이 대규모 실험이 처음 1%의 훈련만 보고도 성공할지, 또는 이 실험의 소규모 버전을 보고 성공할지 예측하는 강력한 기초 과학"을 구축하는 것일 것이다.

- 게다가, 정말 좋은 직관을 넘어서—

탐색 방향에 대해, 제이슨 웨이가 지적했듯이, 수십 개의 하이퍼파라미터와 실험 세부사항에 대해 훌륭한 직관을 갖는 것은 엄청난 수익을 가져온다. 제이슨은 직관에 기반해 첫 시도에서 바로 맞추는 능력을 “요로 실행(yolo runs)”이라고 부른다. (제이슨은 “내가 아는 것은, 이걸 할 수 있는 사람들은 확실히 AI 연구자들보다 10~100배 뛰어나다”라고 말한다.)

컴퓨팅 병목 현상은 백만 배 더 많은 연구자가 있다고 해도 연구 속도가 백만 배 빨라지지 않는다는 것을 의미한다—따라서 단기간 내 지능 폭발은 아니다. 하지만 자동화된 AI 연구자들은 인간 연구자들에 비해 엄청난 이점을

41. 예를 들어 GPT-4 위에 25 OOM의 알고리즘 진보는 명백히 불가능하다: 이는 GPT-4 수준 모델을 단 몇 FLOP만으로 훈련시킬 수 있다는 의미가 되기 때문이다.

가질 것이며, 적어도 컴퓨팅을 10배 이상 더 효율적/효과적으로 사용할 방법을 찾지 못할 이유가 없어 보인다—따라서 알고리즘 진보 속도가 10배 빨라지는 것은 충분히 그럴듯하다.

여기서 잠시, 내가 들은 가장 설득력 있는 반론 중 하나를 내 친구 제임스 브래드버리에게서 인정하고자 한다: 만약 더 많은 ML 연구 노력이 진보를 극적으로 가속화한다면, 왜 현재 수만 명에 달하는 학계 ML 연구 커뮤니티가 최첨단 연구실 진보에 더 크게 기여하지 않는가? (현재로서는, 연구실 내부 팀, 아마도 연구실 전체를 합쳐 천 명 정도가 최첨단 알고리즘 진보의 대부분을 담당하는 것 같다.) 그의 주장은 알고리즘 진보가 컴퓨팅 병목에 걸려 있기 때문이라는 것이다: 학계는 단지 충분한 컴퓨팅 자원을 갖고 있지 않다.

몇 가지 답변:

- 질을 조정해 보면, 학계 연구자는 아마 수만 명이 아니라 수천 명일 것이다 (예: 상위 대학만 고려할 때). 이는 아마 연구실 인원 합계와 크게 다르지 않을 것이다. (그리고 자동화된 AI 연구에서 나올 수백만 명 연구자와는 비교도 안 된다.)
- 학계는 잘못된 주제에 집중한다. 지금까지 매우...

최근(그리고 아마도 지금도?), 학계의 대다수 머신러닝 연구자들은 대형 언어 모델에 전혀 집중하지 않고 있었다. 학계에서 대형 언어 모델을 연구하는 강력한 학자들의 수는 연구소의 연구자들을 합친 것보다 의미 있게 적을 수도 있다?

- 학자들이 LLM 사전학습과 같은 연구를 하더라도, 그들은 최첨단 기술에 접근할 수 없다—연구소 내부의 최전선 모델 훈련에 관한 방대한 세부 지식 축적을 알지 못한다. 그들은 실제로 어떤 문제가 중요한지 알지 못하거나, 기준선이 잘못 조정되어 아무도 실제로 활용할 수 없는 단발성 결과만 낼 수 있다(그래서 그들의 결과가 실제 개선인지 알 수 없다).
- 학자들은 자동화된 AI 연구자들보다 훨씬 못하다: 그들은 인간 속도의 10배 또는 100배로 작업할 수 없고, 지금까지 작성된 모든 ML 논문을 읽고 내재화할 수 없으며, 10년간 모든 코드 라인을 점검하거나, 온보딩 병목 현상을 피하기 위해 스스로 복제할 수도 없다.

학자들에 대한 반대 사례로는 GDM이 OpenAI보다 훨씬 더 많은 실험용 컴퓨팅을 보유했다는 소문이 있지만, 알고리즘 진보 면에서 GDM이 OpenAI를 크게 앞서지 않는 것으로 보인다는 점이 있다.

일반적으로, 자동화된 연구자들은 자신의 강점을 살리고 컴퓨팅 병목 현상을 완화하는 방향으로 다른 연구 방식을 가질 것으로 예상된다. 이 결과가 어떻게 될지 불확실한 것은 합리적이지만, 인간에게 어려운 일이기 때문에 모델들이 컴퓨팅 병목을 극복하지 못할 것이라고 확신하는 것은 비합리적이다.

- 예를 들어, 그들은 초기 단계에서 "작은 규모 실험으로부터 대규모 결과를 예측하는 방법"에 관한 기초 과학을 구축하는 데 많은 노력을 기울일 수 있다. 그리고 인간이 할 수 없는 많은 일을 할 수 있을 것으로 기대한다. 예를 들어, "이 대규모 실험이 성공할지 여부를 실험 초반에 예측하는 것" 같은 일 말이다.

"첫 1%의 훈련". 만약 당신이 매우 뛰어난 자동화 연구자이며 초인적인 직관을 가지고 있다면, 이것은 꽤 실현 가능해 보이며, 이로 인해 많은 컴퓨팅을 절약할 수 있습니다.

- AI 시스템이 AI 연구를 자동화하는 모습을 상상할 때, 저는 이들이 컴퓨팅 병목 현상에 걸리지만, 인간보다 1000배 더 많이(그리고 더 빠르게) 사고하고, 인간보다 더 높은 수준의 품질로 사고함으로써 이를 상당 부분 보

완한다고 봅니다(예: 수백만 개의 실험 결과를 예측하도록 훈련받아 얻은 초인적인 머신러닝 직관 덕분에). 만약 이들이 단순히 엔지니어링보다 사고 능력이 훨씬 떨어지지 않는다면, 이것이 많은 부분을 보완할 수 있다고 생각하며, 이는 학계와는 질적으로 다를 것입니다.

(실험 컴퓨팅 외에도, 결국 몇 달이 걸리는 대규모 훈련 실행이라는 추가 병목 현상이 있습니다. 하지만 이를 절약할 수 있을 것입니다. 지능 폭발의 해 동안 몇 차례만 실행하고, 현재 연구소들이 하는 것보다 더 큰 규모 차수(OOM) 도약을 할 수 있습니다.⁴² 또는 컴퓨팅 효율성 향상 5단계 중 1단계를 "소비"하여 훈련 실행을 몇 달이 아닌 며칠 만에 완료할 수도 있습니다.)

참고로, 저는 이것이 가능할 것이라고 생각하지만, 다소 무섭기도 합니다: 이는 이전 세대보다 다소 나은 일련의 대규모 모델들이 연속적으로 등장하는 대신, 하위 모델 지능이 더 불연속적/비연속적으로 발전할 수 있음을 의미합니다. 우리는 지능 폭발 동안 한두 번의 대규모 실행만 할 수도 있으며, 각 실행마다 소규모에서 발견된 알고리즘 혁신의 여러 규모 차수를 쌓아둘 수 있습니다.

100% 자동화에 대한 상호보완성과 긴 꼬리(선택 사항, 더 깊이)

경제 성장을 가속화하는 AI 자동화에 대한 고전적인 경제학자의 반대는 서로 다른 작업들이 상호보완적이라는 점입니다: 예를 들어, 1800년대 인간 노동의 80%를 자동화했음에도 불구하고 성장 폭발이나 대규모 실업으로 이어지지 않았으며, 나머지 20%가 모든 인간이 수행하는 작업이 되어 병목 현상으로 남았습니다. (여기에서 이 모델을 참조하세요.)

저는 경제학자들의 이 모델이 옳다고 생각합니다. 하지만 중요한 점은 제가 현재 경제 전체가 아니라 현재는 작은 한 부분에 대해서만 이야기하고 있다는 것입니다. 이 기간 동안 사람들은 여전히 정상적으로 이발을 받을 수도 있습니다—로봇 공학이 아직 완성되지 않았거나, 모든 분야에 AI가 아직 도입되지 않았거나, 사회적...

출시 방식이 아직 확정되지 않았을 수도 있지만, 그들은 AI 연구를 수행할 수 있을 것입니다. 앞서 논의한 바와 같이, 현재 AI 발전 경로는 본질적으로 가장 똑똑한 인간만큼 지능적인 원격 근로자를 투입하는 방향으로 나아가고 있습니다. 이 글에서 논의한 바와 같이, AI 연구자의 업무는 완전 자동화가 충분히 가능한 범위 내에 있는 것으로 보입니다.

그럼에도 불구하고, 실제로는 AI 연구자/엔지니어 업무의 100% 완전 자동화에 도달하기까지 다소 긴 시간이 걸릴 것으로 예상합니다. 예를 들어, 처음에는 거의 엔지니어 대체 역할을 하는 시스템이 등장하겠지만, 여전히 일정 수준의 인간 감독이 필요할 수 있습니다.

특히, AI 역량 수준은 분야별로 다소 불균형하고 편차가 클 것으로 예상합니다. 예를 들어, 어떤 분야에서는 최고의 엔지니어보다 더 뛰어난 코더일 수 있지만, 일부 작업이나 기술에서는 약점이 있을 수 있습니다. 가장 약한 분야에서 인간 수준에 도달할 때쯤이면, 코딩처럼 훈련하기 쉬운 분야에서는 이미 상당히 초인적인 능력을 갖추고 있을 것입니다. (이것이 AI가 인간 연구자보다 컴퓨팅을 더 효과적으로 사용할 수 있다고 생각하는 이유 중 하나입니다. 100% 자동화/지능 폭발이 시작될 시점에는 이미 일부 분야에서 인간에 비해 큰 우위를 점하고 있을 것입니다. 이는 초정렬에 중요한 함의를 가지는데, 최초의 자동화된 AI 연구자들을 정렬하기 위해서는 여러 분야에서 의미 있게 초인적인 시스템을 정렬해야 하기 때문입니다.)

42. 저는 이것이 가능성이 높다고 생각하지만, 약간 무섭기도 합니다: 연속적인 큰 모델 시리즈가 아니라, 아마도 지능 폭발 중에 한두 차례의 큰 시도만 해서 각 시도마다 다수 OOM 규모의 알고리즘적 돌파구를 축적할 수 있음을 의미할 수 있습니다.

하지만 이 단계가 몇 년 이상 지속되지는 않을 것으로 봅니다. AI 발전 속도를 고려할 때, 마지막 단계를 완성하는 데는 모델의 명백한 한계를 제거하는 추가적인 ‘제약 해소’나 다음 세대 모델이 필요할 뿐일 것입니다.

전반적으로, 이 과정은 급격한 도약을 다소 완화할 수 있습니다. 2027년 범용 인공지능(AGI) → 2028년 초지능으로 바로 이어지기보다는 다음과 같은 모습일 수 있습니다:

- 2026/27년: 원시 자동화 엔지니어, 그러나 일부 약점 존재

기타 분야. 이미 작업 속도가 1.5배에서 2배 빨라짐; 진전이 점차 가속화되기 시작함.

- 2027/28: 프로토 자동 연구자, 90% 이상 자동화 가능. 일부 남아있는 인간 병목 현상과 거대한 자동 연구자 조직을 조율하는 데 발생하는 문제들이 해결되어야 하지만, 이미 진전 속도를 3배 이상 가속화함. 이는 빠르게 나머지 필요한 “제약 해소”를 수행하여 100% 자동화에 이르게 함.

- 2028/29: 진전 속도 10배 이상 → 초지능.

이것도 여전히 매우 빠른 속도이다...

알고리즘 진보의 근본적 한계 (선택 사항, 더 깊이)

알고리즘 진보가 물리적으로 가능한 최대치에는 실제 한계가 있을 가능성이 높다. (예를 들어, 25 규모 차수(OOM)의 알고리즘 진보는 불가능해 보이는데, 이는 GPT-4 수준 시스템을 약 10 FLOPs 미만으로 훈련할 수 있다는 것을 의미하기 때문이다.⁴³) 하지만 약 5 OOM 정도는 충분히 가능성 있는 범위로 보인다; 다시 말해, 이는 단지 알고리즘 효율성의 추세가 10년 더 지속되어야 한다는 것을 의미한다 (제약 해소를 인한 알고리즘 향상은 포함하지 않음).

직관적으로, 가장 큰 돌파구들이 얼마나 단순한지, 그리고 현재의 아키텍처와 훈련 기법들이 얼마나 원시적이고 명백히 제약되어 있는지를 고려할 때, 아직 저절로 얻을 수 있는 성과를 모두 소진했다고 보이지 않는다. 예를 들어, 나는 AI 시스템이 ‘사고 과정을 소리 내어 생각하는’ 체인 오브 쏫트(chain-of-thought)를 통해 범용 인공지능(AGI)에 도달할 가능성이 꽤 있다고 생각한다. 하지만 분명히 이것이 가장 효율적인 방법은 아니며, 내부 상태/순환 등으로 이 추론을 수행하는 무언가가 훨씬 더 효율적일 것이다. 또는 적응형 컴퓨팅을 고려해 보자: Llama 3는 복잡한 질문에 대한 답변을 예측하는 데 쓰는 컴퓨팅과 ‘and’ 토큰을 예측하는 데 쓰는 컴퓨팅이 동일인데, 이는 명백히 비효율적이다. 우리는 작은 조정만으로도 엄청난 규모 차수(OOM)의 알고리즘 이득을 얻고 있으며, 훨씬 더 효율적인 수십 개의 분야가 있다.

효율적인 아키텍처와 훈련 절차도 아마 발견될 수 있을 것이다.

생물학적 참조도 큰 여지가 있음을 시사한다. 예를 들어, 인간의 지능 범위는 매우 넓으며, 아키텍처에 약간의 조정만으로도 가능하다. 인간은 다른 동물들과 비슷한 수의 뉴런을 가지고 있지만, 인간이 그 동물들보다 훨씬 더 똑똑하다. 그리고 현재 AI 모델들은 인간 두뇌의 효율성에서 아직도 여러 규모 차수(OOM) 차이가 있다; 인간은 AI 모델들이 학습하는 것보다 훨씬 적은 데이터(따라서 훨씬 적은 “컴퓨팅”)로 학습할 수 있는데, 이는 우리의 알고리즘과 아키텍처에 큰 여지가 있음을 시사한다.

아이디어는 찾기 어려워지고 수확 체감이 발생한다 (선택 사항, 더 깊이)

43. 물론 현재 아키텍처로는 25 OOM 더 많은 하드웨어가 필요한 결과를 얻을 수도 있다!

낮은 열매를 따면서 아이디어를 찾기 점점 어려워진다. 이는 기술 진보의 어떤 영역에서도 사실이다. 본질적으로, 우리는 로그-로그 곡선에서 직선을 본다: 로그(진보)는 로그(누적 연구 노력)의 함수이다. 추가적인 진보의 각 규모 차수(OOM)는 이전 규모 차수보다 더 많은 연구 노력을 투입해야 한다.

이것은 지능 폭발에 대한 두 가지 반론으로 이어진다:

1. 자동화된 AI 연구는 진보를 극적으로 가속화하기보다는 단지 진보를 유지하는 데 필요한 수준에 그칠 것이다.
2. 순수 알고리즘적 지능 폭발은 지속되지 않거나 알고리즘 진보를 찾기 어려워지고 한계 수확 체감에 부딪혀 빠르게 사그라들 것이다.

나는 과거 경제학 연구를 할 때 이러한 유형의 모델에 대해 많은 시간을 생각했다. (특히, 반내생 성장 이론은 기술 진보의 표준 모델로, 증가하는 연구 노력과 아이디어 찾기 어려움이라는 두 경쟁 역학을 포착한다.)

요컨대, 이러한 반대 의견들의 근본적인 모델은 타당하다고 생각하지만, 그것이 어떻게 전개될지는 경험적인 문제이며—그들이 경험적 분석을 잘못했다고 생각한다.

핵심 질문은 본질적으로 다음과 같다: 진보가 10배 이루어질 때마다, 추가 진보는 10배 더 어려워지는가 아니면 덜 어려워지는가? 경제학 문헌에서 이와 유사하게 수행되는 간단한 계산이 이를 제한하는 데 도움이 된다.

- 알고리즘 진보의 연간 약 0.5 규모 차수(OOM) 추세를 진지하게 받아들인다면, 4년 내에 100배의 진보를 의미한다.
- 그러나 특정 선도 AI 연구소에서의 질 보정된 인력/연구 노력은 4년 동안 확실히 100배 미만으로 증가했다. 아마도 10배 증가했을 수 있다(해당 연구소에서 관련 업무를 하는 인원이 수십 명에서 수백 명으로), 하지만 질 보정을 감안하면 그마저도 불확실하다.
- 그럼에도 불구하고, 알고리즘 진보는 지속되는 것으로 보인다.

따라서 반대 의견 1에 대한 답변으로, 연구 노력의 약 백만 배 증가는 단순히 진보를 유지하는 데 필요한 것보다 훨씬 더 큰 증가일 것임을 지적할 수 있다. 아마도 4년 내에 한 연구소에서 관련 연구를 수행하는 연구자가 수천 명 정도 필요할 수 있는데, 1억 명의 Alec Radford가 있다면 여전히 엄청난 증가이며 분명히 대규모 가속을 초래할 것이다. 자동화된 연구가 기존 진보 속도를 유지하는 데 딱 맞는 수준일 것이라는 가정은 매우 기이한 ‘칼날 끝’ 가정이다. (그리고 이는 인간 연구자보다 10배 빠른 사고 속도와 AI 시스템이 가질 다른 모든 이점을 고려하지 않은 것이다.)

반대 의견 2에 대한 답변으로, 두 가지를 지적할 수 있다:

- 첫째, 위에서 언급한 수학적 조건이다. 우리의 간단한 계산에 따르면, 알고리즘 진보가 100배 이루어지는 동안 질 보정된 연구 노력은 100배 미만으로 증가해야 했으므로, 이는 상당히 강력하게 보인다.

수익 곡선의 형태는 자가 지속적 진보에 유리하게 나타난다.⁴⁴

44. 진보 $\times \rightarrow$ 자동화 연구 역량 $\times 100$, 하지만 다음 100배를 지속하기 위해 10배의 연구 역량만 필요하다면 수익이 충분하여 폭발적 진보가 유지될 수 있습니다.

Tom Davidson과 Carl Shulman도 성장 모델링 틀에서 이 경험적 분석을 수행하여 유사한 결론에 도달했다. Epoch AI도 최근 경험적 연구를 수행했으며, 알고리즘 연구개발에 대한 경험적 수익이 폭발적 성장을 지지한다는 결론과 그 함의에 대한 유용한 설명을 내놓았다.

초지능의 힘

이러한 주장 중 가장 강력한 형태에 동의하든—1년 미만의 지능 폭발이 오든, 몇 년이 걸리든—명확한 것은: 우리는 초지능의 가능성에 맞서야 한다는 점이다.

이번 10년 말까지 우리가 갖게 될 AI 시스템은 상상할 수 없을 정도로 강력할 것이다.

- 물론, 이들은 양적으로 인간을 초월할 것이다. 우리의 함대에서

수십억 개의 GPU를 운용할 수 있게 되면, 우리는 수십억 개의 GPU로 구성된 문명을 운영할 수 있게 될 것이며, 이들은 인간보다 규모 차수(OOM)로 훨씬 빠르게 “생각”할 수 있을 것이다. 이들은 어떤 분야든 빠르게 마스터할 수 있고, 수조 줄의 코드를 작성하며, 지금까지 작성된 모든 과학 분야의 연구 논문을 읽을 수 있을 것이다(완벽하게 학제간 지식을 갖추게 될 것이다!). 그리고 당신이 한 논문의 초록을 넘기기도 전에 새로운 논문을 작성할 것이며, 그 복제본 각각의 병렬 경험으로부터 학습하고, 몇 주 만에 수십억 년에 해당하는 인간과 동등한 경험을 새로운 혁신에 대해 쌓을 것이다. 100%의 시간 동안 최고 에너지와 집중력으로 작업하며, 뒤처지는 팀원 때문에 속도가 느려지지 않을 것이다 등등.

- 더 중요하지만 상상하기 어려운 점은, 이들이 질적으로 초인적일 것이라는 점이다. 좁은 예로, 대규모 강화학습(RL) 실행은 인간의 이해를 넘어서는 완전히 새롭고 창의적인 행동을 만들어냈는데, 예를 들어 AlphaGo가 이세돌과의 대국에서 보여준 유명한 37번째 수가 있다. 초지능은 여러 분야에서 이와 같을 것이다. 인간이 알아차리기에는 너무 미묘한 인간 코드의 취약점을 찾아내고, 모델이 수십 년을 설명하려 해도 인간이 이해할 수 없을 정도로 복잡한 코드를 생성할 것이다. 인간이 수십 년간 막혀 있을 극도로 어려운 과학적·기술적 문제들이 그들에게는 너무나 명백하게 보일 것이다. 우리는 고등학생이 뉴턴 역학에 갇혀 있을 때 그들이 양자역학을 탐구하는 것과 같을 것이다.

이것이 얼마나 놀라운 일인지 예를 들자면, 유튜브에서 비디오 게임 스피드런 영상을 보라. 예를 들어 마인크래프트를 20초 만에 클리어하는 영상이 있다. (이 영상에서 무슨 일이 일어나고 있는지 전혀 모르겠다면, 당신은 좋은 동료가 있다; 대부분의 일반 마인크래프트 플레이어들도 무슨 일이 일어나고 있는지 거의 알지 못한다.)

이제 이것이 과학, 기술, 경제의 모든 분야에 적용된다고 상상해 보라. 물론 여기에는 매우 큰 오차 범위가 있다. 그럼에도 불구하고, 이것은 실제로 일어나고 있으며, 얼마나 중대한 결과를 초래할지 고려하는 것이 중요하다.

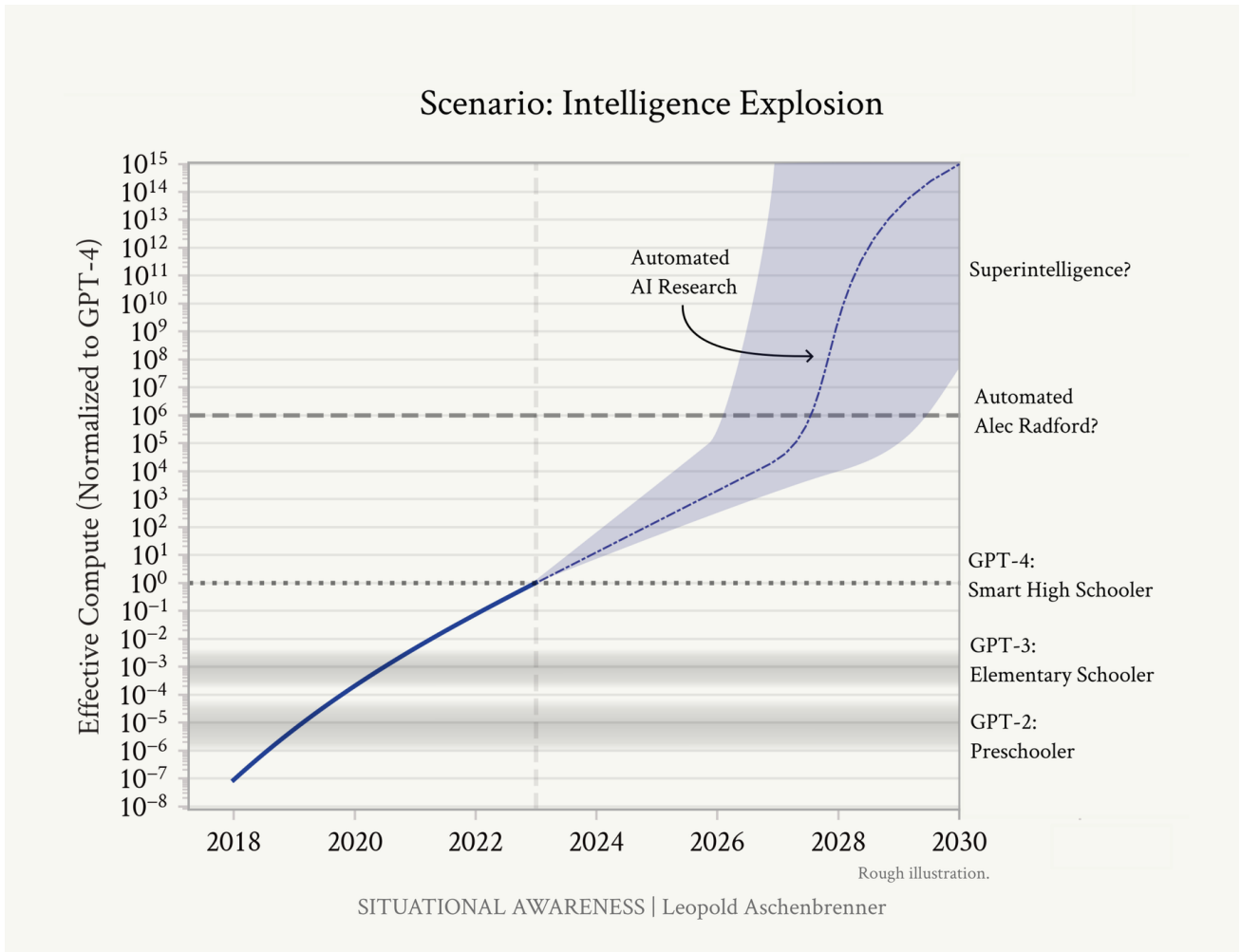


그림 25: 여기 서 있으면 어떤 느낌일까? Wait But Why/팀 어반(Tim Urban)에서 발췌한 삽화.

지능 폭발에서는 초기의 폭발적 진보가 자동화된 AI 연구라는 좁은 영역에만 국한되었다. 범용 초지능이 등장하고 수십억의 (이제는 초지능화된) 에이전트를 다양한 분야의 연구개발에 투입하면, 폭발적 진보는 다음과 같이 확장될 것으로 예상된다:

- AI 역량 폭발. 초기의 범용 인공지능(AGI)은 일부 다른 영역(단지 AI 연구 영역뿐만 아니라)에서 작업을 완전히 자동화하는 데 제한이 있었을 수 있다; 자동화된 AI 연구가 이를 빠르게 해결하여 모든 인지 작업의 자동화를 가능하게 할 것이다.
- 로봇틱스 문제 해결. 초지능은 오래도록 순수 인지 기능에만 머무르지 않을 것이다. 로봇틱스를 잘 작동시키는 것은 주로 하드웨어 문제가 아니라 머신러닝 알고리즘 문제이며, 우리의 자동화된 AI 연구자들이 이를 해결할 가능성이 높다(아래에서 더 설명!). 공장은 인간이 운영하는 단계에서 AI가 인간의 물리적 노동을 지시하는 단계로, 곧 로봇 무리가 완전히 운영하는 단계로 전환될 것이다.
- 과학 및 기술 진보의 극적인 가속화. 아인슈타인 한 명이 신경과학을 발전시키고 반도체 산업을 구축할 수는 없었지만, 수십억의 초지능 자동화 과학자, 엔지니어, 기술자, 로봇이 있다면

기술자들(로봇이 인간 속도의 10배 이상으로 움직임!)⁴⁶은 수년 내에 여러 분야에서 놀라운 발전을 이룰 것이다. (AI 주도의 연구개발이 어떻게 보일지 시각화한 짧은 이야기 하나가 있다.) 수십억의 초지능들은 인간 연구자들

이 다음 세기 동안 할 연구개발 노력을 수년 내로 압축할 수 있을 것이다. 20세기의 기술 진보가 10년도 채 안 되는 기간에 압축된다면 어떻게 상상해 보라. 우리는 비행이 환상으로 여겨지던 시절에서 비행기, 달 착륙, 대륙간탄도미사일(ICBM)까지 수년 만에 도달했을 것이다. 이것이 내가 2030년대 과학과 기술 전반에서 기대하는 모습이다.

- 산업 및 경제 폭발. 극도로 가속화된 기술 진보와 모든 인간 노동의 자동화 능력이 결합되면 경제 성장이 극적으로 가속화될 수 있다(예: 네바다 사막 전체를 빠르게 덮는 자기 복제 로봇 공장⁴⁷). 성장률 증가는 단순히 연 2%에서 2.5%로의 증가가 아니라, 매우 느린 성장에서 산업혁명과 함께 연 몇 퍼센트 성장으로의 역사적 단계 변화에 비견될 근본적인 성장 체제의 변화일 것이다. 우리는 연 30% 이상의 경제 성장률, 심지어 연간 여러 차례의 두 배 증가를 볼 수도 있다. 이는 경제학자들의 경제 성장 모델에서 비교적 직관적으로 도출된다. 물론, 사회적 마찰로 인해 지연될 수도 있다; 복잡한 규제가 AI 시스템이 훨씬 더 잘할 수 있는 직업임에도 불구하고 변호사와 의사는 여전히 인간이어야 한다고 보장할 수 있고; 사회가 변화 속도를 저항하면서 급속히 확장되는 로봇 공장에 모래를 던질 수도 있으며; 인간 보모를 유지하고 싶어 할 수도 있다; 이 모든 것이 전체 GDP 통계의 성장을 늦출 것이다. 그럼에도 불구하고, 인간이 만든 장벽(예: 군사 생산을 위해 경쟁이 우리를 강제할 수도 있음)을 제거하는 어떤 영역에서는 산업 폭발을 보게 될 것이다.

표 3: 성장 체제의 전환 성장 모드 지배하기 시작한 시기 전 세계 경제의 문명이 사냥에서 농업으로, 두 배 증가 시간(년) 과학과 상업의 번영, 산업으로 전환하면서 전례 없는 일이 아니다. 사냥 기원전 2,000,000년 230,000 전 세계 경제 성장 속도가 가속화되었다. 초지능은 성장 모드의 또 다른 전환을 농업 기원전 4700년 860 촉발할 수 있다. 로빈 핸슨의 과학/상업 서기 1730년 58 “장기 성장: 지수적 모드의 연속”에 근거한다. 산업 서기 1903년 15 초지능? 서기 2030년? ???

- 결정적이고 압도적인 군사적 우위를 제공한다. 초기 인지 초지능만으로도 충분할 수 있다; 아마도 어떤 초인적 해킹 계획이 적군 군대를 무력화할 수 있을 것이다. 어쨌든 군사력과 기술 진보는 역사적으로 밀접하게 연관되어 왔으며, 매우 빠른 기술 발전과 함께 군사 혁명도 뒤따를 것이다. 드론 떼와 로봇 군대가 큰 이슈가 되겠지만, 그것들은 시작에 불과하다; 새로운 유형의 무기들, 신종 대량살상무기부터 무적의 레이저 기반 미사일 방어체계, 그리고 아직 상상할 수 없는 것들이 등장할 것이다. 초지능 이전의 무기고와 비교하면, 21세기 군대가 19세기 말 말과 총검을 든 여단과 싸우는 것과 같다. (초지능이 어떻게 결정적인 군사 우위를 가져올 수 있는지 후속 글에서 다룰 것이다.)

- 미국 정부를 전복할 수 있다. 초지능을 통제하는 자는 초지능 이전 세력으로부터 권력을 탈취할 만큼 충분한 힘을 가질 가능성이 크다. 로봇이 없어도, 초지능 문명은 방어되지 않은 군사, 선거, 텔레비전 등 시스템을 해킹할 수 있고, 장군과 유권자를 교묘히 설득하며, 국가를 경제적으로 능가하고, 새로운 합성 생물무기를 설계한 뒤 비트코인으로 인간에게 합성을 의뢰하는 등 다양한 수단을 사용할 수 있다. 1500년대 초, 코르테스와 약 500명의 스페인인이 수백만 명의 아즈텍 제국을 정복했고, 피사로와 약 300명의 스페인인이 수백만 명의 잉카 제국을

46. 실제 세계에서 물리적 연구개발을 수행하는 10배 속도의 로봇은 “느린 버전”이다; 실제로 초지능들은 AlphaFold나 제조 “디지털 트윈”처럼 가능한 한 많은 연구개발을 시뮬레이션에서 수행하려 할 것이다.

47. “팩토리오 세계”—공장을 지어 더 많은 공장을 생산하고, 더 많은 공장을 생산하는 공장을 만들어 결국 행성 전체가 공장으로 빠르게 덮이는 것이 왜 오늘날 불가능한가? 노동력이 제한되기 때문이다—자본(공장, 도구 등)을 축적할 수는 있지만, 고정된 노동력에 의해 수익 체감이 발생한다. 로봇과 AI 시스템이 노동을 완전히 자동화할 수 있다면, 그 제약이 제거된다; 로봇 공장은 거의 무제한적으로 더 많은 로봇 공장을 생산할 수 있어 산업 폭발로 이어질 수 있다. 이에 관한 경제 성장 모델은 여기에서 더 볼 수 있다.

정복했으며, 알폰소와 약 1000명의 포르투갈인이 인도양을 정복했다. 그들은 신과 같은 힘을 가진 것은 아니었지만, 구세계의 기술 우위와 전략적 우위가 있었다.

지능과 외교적 교활함이 완전히 결정적인 우위를 가져왔다. 초지능도 비슷하게 보일 수 있다.

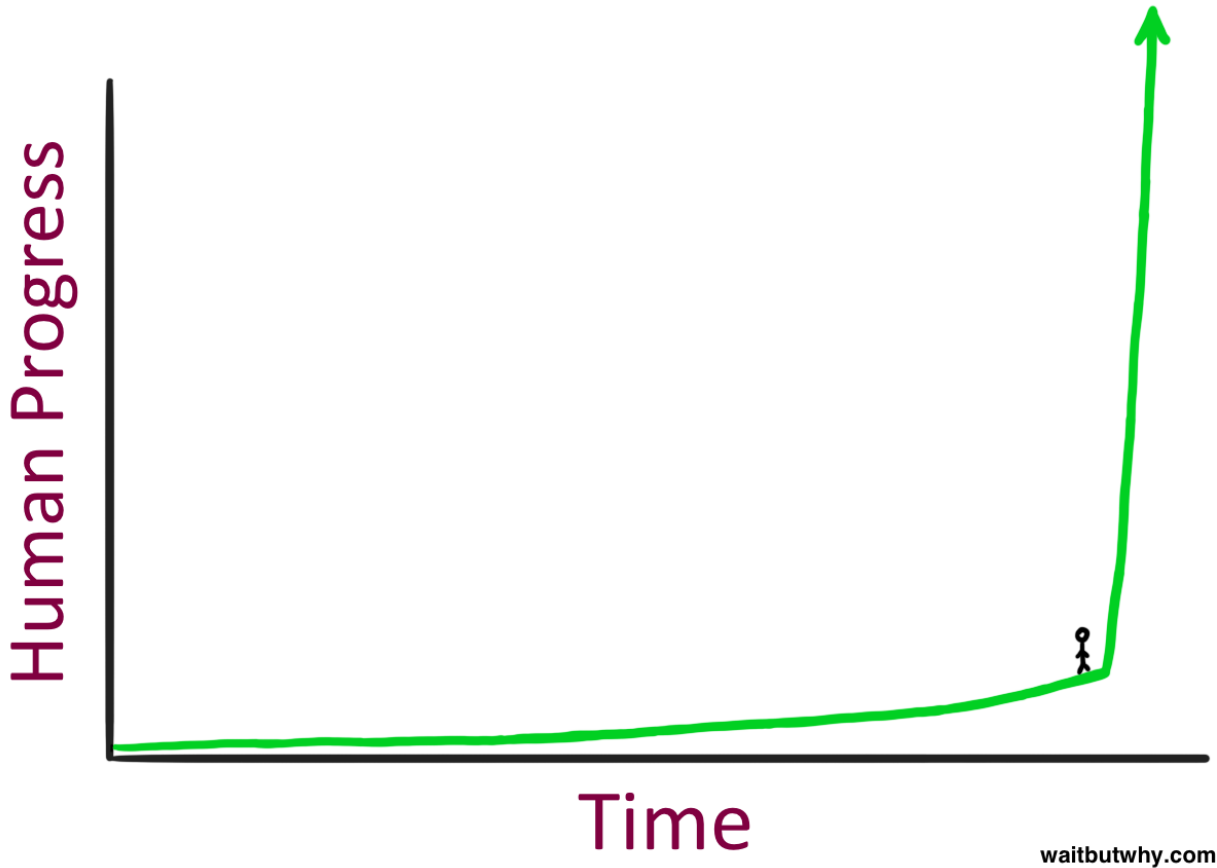


그림 26: 폭발적 성장은 AI 연구개발의 로봇. 여기서 제기되는 일반적인 반론 중 하나는, 설령 AI가 인지 작업을 수행할 수 있다 하더라도, 로봇 분야가 크게 뒤쳐져 있어 실제 세계에 미치는 영향에 제동이 걸릴 것이라는 점이다. 좁은 영역에서 시작된다; 초지능을 다른 분야의 연구개발에 적용할수록 폭발적 성장은 더 넓어질 것이다.

나는 한때 이 의견에 공감했지만, 이제는 로봇이 장애물이 되지 않을 것이라고 확신하게 되었다. 수년간 사람들은 로봇이 하드웨어 문제라고 주장했지만, 로봇 하드웨어는 이미 해결 단계에 접어들고 있다.

점점 더 명확해지는 것은 로봇이 기계학습(ML) 알고리즘 문제라는 점이다. 대형 언어 모델(LLM)은 부트스트랩이 훨씬 쉬웠다: 인터넷 전체를 사전학습 데이터로 활용할 수 있었기 때문이다. 로봇 동작에 대해 이와 유사한 대규모 데이터셋은 없으며, 따라서 멀티모달 모델을 기반으로 삼고 합성 데이터/시뮬레이션/영리한 강화학습(RL) 등을 활용하는 등 더 정교한 접근법이 필요하다.

지금 이 문제를 해결하기 위해 엄청난 에너지가 집중되고 있다. 하지만 AGI 이전에 이 문제를 해결하지 못하더라도, 수억 대의 AGI/초지능들이 놀라운 AI 연구자가 될 것이며(이 글의 핵심 주장이다!), 그들이 놀라운 로봇을 작동시키는 머신러닝 문제를 반드시 해결할 가능성이 매우 높다.

따라서 로봇이 몇 년 정도 지연을 초래할 가능성은 있지만(머신러닝 문제 해결, 시뮬레이션보다 근본적으로 느린 물리적 세계에서의 테스트, 로봇이 스스로 공장을 짓기 전에 초기 로봇 생산 확대 등), 그 이상은 아닐 것이라고 생각한다.

이 모든 일이 2030년대에 어떻게 전개될지는 예측하기 어렵다(그리고 그것은 또 다른 이야기다). 하지만 적어도 한 가지는 분명하다: 우리는 인류가 지금까지 직면한 것 중 가장 극단적인 상황에 빠르게 몰입할 것이다.

인간 수준의 AI 시스템, 즉 범용 인공지능(AGI)은 그 자체로도 매우 중대한 의미를 가지겠지만, 어떤 면에서는 우리가 이미 알고 있는 것의 더 효율적인 버전에 불과할 것이다. 그러나 매우 그럴듯하게도, 단 1년 내에 우리는 훨씬 더 이질적인 시스템으로 전환할 것이며, 그 시스템의 이해력과 능력—즉 순수한 힘—은 인류 전체를 합친 것조차 능가할 것이다. 우리는 이 급격한 전환 과정에서 AI 시스템에 신뢰를 맡길 수밖에 없게 되면서 통제력을 잃을 가능성이 실제로 존재한다.

더 일반적으로, 모든 일이 믿을 수 없을 정도로 빠르게 진행될 것이다. 그리고 세상은 미쳐 돌아가기 시작할 것이다. 만약 우리가 20세기의 지정학적 격변과 인위적 위험들을 단 몇 년 만에 겪었다면; 그것이 바로 초지능 이후에 우리가 예상해야 할 상황이다. 그 끝에는 초지능 AI 시스템이 우리의 군사와 경제를 운영하고 있을 것이다. 이 모든 광란 속에서 우리는 올바른 결정을 내릴 시간이 극히 부족할 것이다. 도전 과제는 엄청날 것이다. 무사히 이 상황을 헤쳐 나가려면 우리가 가진 모든 것을 쏟아부어야 할 것이다.

지능 폭발과 초지능 직후의 시기는 인류 역사상 가장 불안정하고 긴장되며 위험하고 격렬한 시기 중 하나가 될 것이다.

그리고 아마도 이 10년이 끝날 무렵 우리는 그 한가운데에 있을 것이다.

지능 폭발—초지능의 출현—가능성에 직면하는 것은 종종 핵 연쇄 반응 가능성에 관한 초기 논쟁을 떠올리게 한다—그리고 그것이 가능하게 할 원자폭탄. H.G. 웰스는 1914년 소설에서 원자폭탄을 예견했다. 1933년 지라드는 연쇄 반응 아이디어를 처음 구상했을 때 아무도 믿지 않았고, 그것은 순전히 이론에 불과했다. 1938년 핵분열이 실험적으로 발견되자 지라드는 다시 크게 놀라 비밀 유지의 필요성을 강력히 주장했고, 몇몇 사람들은 폭탄 가능성에 눈을 뜨기 시작했다. 아인슈타인은 연쇄 반응 가능성을 고려하지 않았지만, 지라드가 그에게 직면했을 때 그는 그 함의를 빠르게 이해하고 필요한 모든 일을 할 준비가 되어 있었다; 그는 경고음을 울리는 것을 두려워하지 않았고 어리석게 보이는 것도 개의치 않았다. 그러나 페르미, 보어, 그리고 대부분의 과학자들은 폭탄 가능성의 비범한 함의를 진지하게 받아들이기보다는 “보수적”인 태도로 축소하는 것이 낫다고 생각했다. 비밀 유지(독일인과의 기술 공유를 피하기 위해)와 기타 전면적인 노력들은 그들에게 터무니없어 보였다. 연쇄 반응은 너무 터무니없는 소리였다. (결국 폭탄이 현실이 되기까지 반십 년밖에 걸리지 않았음에도 불구하고.)

우리는 다시 한 번 연쇄 반응 가능성에 직면해야 한다. 아마도 당신에게는 추측처럼 들릴지 모른다. 그러나 AI 연구소의 고위 과학자들 사이에서는 급격한 지능 폭발이 매우 그럴듯하다고 보는 이들이 많다. 그들은 그것을 볼 수 있다. 초지능은 가능하다.

III. 도전 과제들

IIIa. 조 단위 달러 클러스터 경쟁

가장 놀라운 기술-자본 가속화가 시작되었다. AI 수익이 급격히 증가함에 따라, 수조 달러가 10년 말까지 GPU, 데이터센터, 전력 인프라 구축에 투입될 것이다. 산업 동원은 미국 전력 생산을 수십 퍼센트 이상 증가시키는 것을 포함하여 매우 강력할 것이다.

보시다시피, 나는 이 모든 나라를 공장으로 만들지 않고는 불가능하다고 말했지. 당신은 바로 그렇게 해냈어. 닐스 보어 (1944년 맨해튼 프로젝트 규모를 알게 된 후 에드워드 텔러에게)

AGI 경쟁은 단지 코드와 노트북 뒤에서 벌어지는 것이 아니다—미국의 산업력을 동원하는 경쟁이 될 것이다. 최근 실리콘밸리에서 나온 어떤 것과도 달리, AI는 거대한 산업 공정이다: 각 새로운 모델은 거대한 새로운 클러스터, 곧 거대한 새로운 발전소, 그리고 결국 거대한 새로운 칩 제조 공장을 필요로 한다. 관련 투자는 엄청나다. 그러나 이 모든 것은 이미 비밀리에 진행 중이다.

이 장에서는 이 상황이 무엇을 의미하는지 감을 잡을 수 있도록 수치를 안내할 것이다:

- AI 제품에서 나오는 수익이 급격히 증가함에 따라—2026년경 구글이나 마이크로소프트 같은 기업들이 강력하지만 범용 인공지능 이전 단계의 시스템으로 연간 1,000억 달러 매출에 도달할 가능성이 있다—

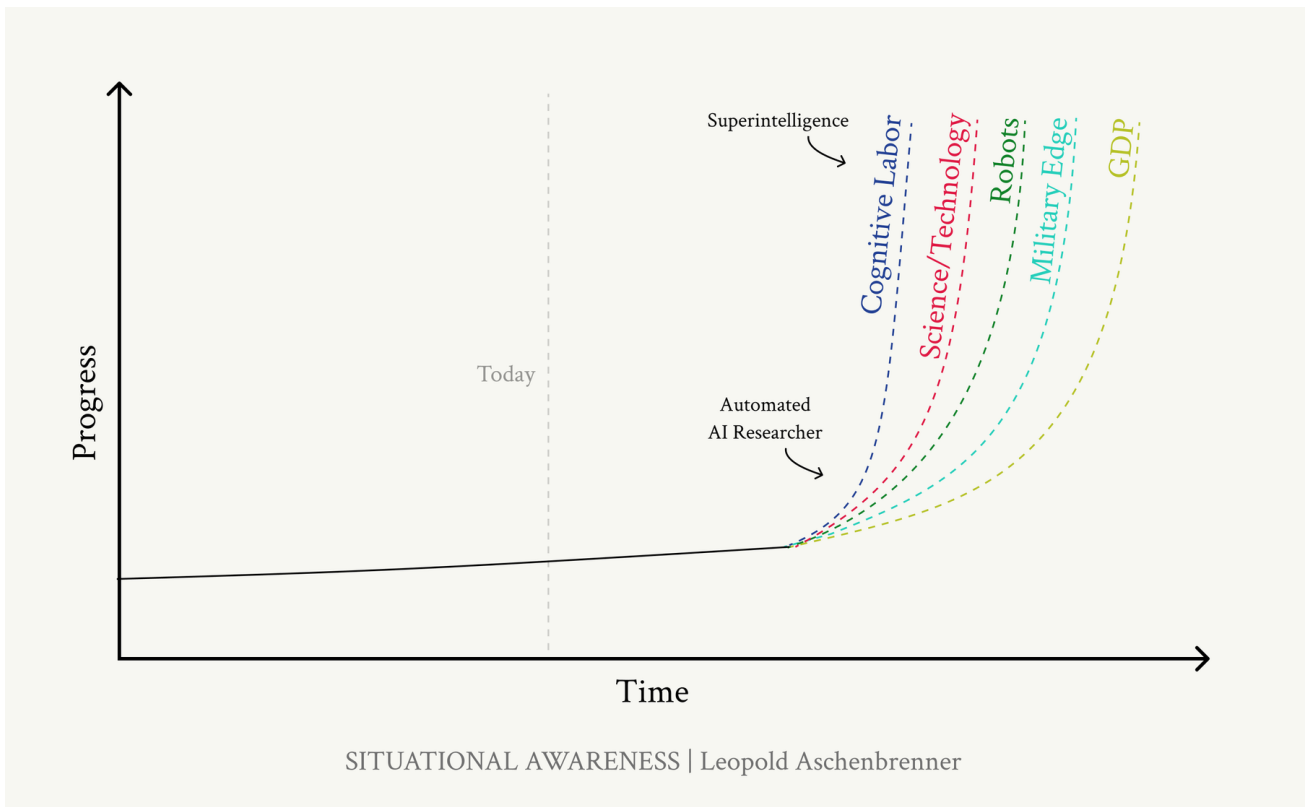


그림 27:조 단위 클러스터. 출처: DALLE. 이는 점점 더 큰 자본 동원을 촉진할 것이며, 총 AI 투자는 2027년까지 연간 1조 달러를 넘을 수 있다.

- 2028년까지 개별 훈련 클러스터 비용이 수천억 달러에 이를 경로에 있다—이 클러스터들은 미국의 소형/중형 주와 맞먹는 전력을 필요로 하며 국제우주정거장보다 더 비싸다.

- 10년 말까지는 1조 달러 이상의 개별 훈련 클러스터로 향하고 있으며, 미국 전력 생산량의 20% 이상에 해당하는 전력을 필요로 한다. 수조 달러의 자본 지출이 매년 수억 개의 GPU를 생산할 것이다.

엔비디아는 데이터센터 매출이 연간 약 140억 달러에서 약 900억 달러로 폭발적으로 증가하며 세계를 놀라게 했다. 하지만 이것은 아직 시작에 불과하다.

학습 컴퓨팅

앞서 우리는 AI 학습 컴퓨팅이 연간 약 0.5 규모 차수(OOM)⁴⁸씩 증가하는 추세를 발견했다. 만약 이 추세가 앞으로 10년 동안 계속된다면, 가장 큰 학습 클러스터는 어떻게 될까?

연도 규모 차수(OOM) H100 동등 수 비용 전력 전력 참조 대상 2022 ~GPT-4 클러스터 ~1만 대 약 5억 달러 약 10MW 약 1만 가구 평균 전력 소비량 ~2024 +1 OOM ~10만 대 수십억 달러 약 100MW 약 10만 가구 ~2026 +2 OOM ~100만 대 수십억 달러 약 1GW 후버 댐 또는 대형 원자력 발전소 ~2028 +3 OOM ~1,000만 대 수백억 달러 약 10GW 미국 소규모/중간 규모 주 ~2030 +4 OOM ~1억 대 1조 달러 이상 약 100GW 미국 전력 생산량의 20% 이상

표 4: 가장 큰 학습 클러스터의 확장, 대략적인 계산. 계산 세부사항은 부록 참조.

믿기 어려울 수 있지만, 실제로 이런 일이 벌어지고 있다. 저커버그는 35만 대의 H100을 구매했고, 아마존은 원자력 발전소 옆에 1GW 데이터센터 캠퍼스를 지었다. 루머에 따르면 쿠웨이트에 1GW, 140만 대 H100 동등 클러스터(약 2026년 클러스터)가 건설 중이다. 언론 보도에 따르면 마이크로소프트와 오픈AI는 2028년 완공 예정인 1,000억 달러 규모 클러스터를 준비 중이라고 한다(국제우주정거장 비용과 비슷한 규모!). 그리고 매 세대 모델이 세상을 놀라게 할 때마다 가속화는 더욱 가속될 수 있다.

아마도 가장 놀라운 점은, 적어도 학습 클러스터에 관해서는 지출 의지가 현재 제약 조건이 아닌 것처럼 보인다는 점이다. 실제 제약은 인프라를 찾는 것이다: “10GW를 어디서 구하지?”(1,000억 달러 이상, 2028년 추세 클러스터의 전력)라는 질문은 샌프란시스코에서 자주 나오는 화제다. 컴퓨팅 담당자들이 가장 고민하는 것은 전력 확보, 부지, 허가, 데이터센터 건설이다.⁴⁹ GPU를 받기 위해 1년을 기다리는 것보다 이 기간이 훨씬 더 길 수 있다.

1조 달러 클러스터—GPT-4 클러스터보다 +4 규모 차수(OOM), 현재 추세로 2030년경의 훈련 클러스터—는 진정으로 엄청난 노력이 될 것이다. 그것이 요구하는 100GW의 전력은 미국 전력 생산량의 20% 이상에 해당한다; 단순히 GPU가 있는 창고가 아니라 수백 개의 발전소를 상상해 보라. 아마도 국가 컨소시엄이 필요할 것이다.

(참고로 나는 AGI에 대해 약 1,000억 달러 규모 클러스터면 충분할 가능성이 꽤 높다고 생각한다. 1조 달러 클러스터는 초지능을 훈련하고 운영하는 데 쓰이거나, AGI가 예상보다 더 어려울 경우 AGI에 사용할 수도 있다. 어쨌든, AGI 이후의 세계에서는 가장 많은 컴퓨팅 자원을 보유하는 것이 여전히 매우 중요할 것이다.)

전체 컴퓨팅

48. 앞서 언급했듯 OOM = 자릿수 단위, 10배 = 1 자릿수입니다.

49. 한 가지 주요 불확실성은 분산 학습이 어떻게 이루어질지다—만약 한 곳에서 그만큼의 전력이 필요하지 않고 100개 장소에 나누어 분산할 수 있다면 훨씬 쉬워질 것이다.

위 수치는 가장 큰 훈련 클러스터에 대한 대략적인 숫자일 뿐이다. 전체 투자는 훨씬 더 클 가능성이 높다: GPU의 상당 부분은 추론⁵⁰(제품용 AI 시스템을 실제로 실행하는 GPU)에 사용될 것이며, 여러 플레이어가 거대한 클러스터를 보유한 경쟁 구도가 될 수 있다.

내 대략적인 추정으로는 2024년에는 이미 1,000억~2,000억 달러 규모의 AI 투자가 이루어질 것이다:

- 엔비디아 데이터센터 매출은 곧 분기 약 250억 달러의 실행 속도에 도달할 것이며, 즉 엔비디아 단독으로 약 1,000억 달러의 자본 지출이 흐를 것이다. 물론 엔비디아만 있는 것은 아니며(구글의 TPU도 훌륭하다!), 데이터센터 자본 지출의 거의 절반은 칩 이외의 부분(부지, 건물, 냉각, 전력 등)에 쓰인다.⁵¹
- 대형 기술 기업들은 자본 지출 규모를 극적으로 늘리고 있다: 마이크로소프트와 구글은 각각 500억 달러 이상⁵², AWS와 메타는 400억 달러 이상을 올해 자본 지출할 것이다. 이 모두가 AI에만 쓰이는 것은 아니지만, AI 붐으로 인해 이들의 자본 지출은 연간 500억~1,000억 달러가 증가했으며, 그럼에도 불구하고 AI에 더 많은 지출을 하기 위해 다른 자본 지출을 줄이고 있다.

더욱이, 다른 클라우드 제공업체, 기업들(예: 테슬라는 올해 AI에 100억 달러를 투자 중), 그리고 국가들도 AI에 투자하고 있다.

50. 참고로, 예를 들어 저크(Zuck)를 보면, 그의 H100 GPU 약 4만 5천 개만이 가장 큰 훈련 클러스터에 있고, 나머지 35만 개는 추론에 사용된다. 메타는 다른 플레이어보다 추론 수요가 더 많을 가능성이 크며, 다른 플레이어들은 지금까지 고객 수가 적어 추론 수요가 적다. 하지만 다른 모든 AI 제품이 확장됨에 따라 추론이 GPU의 강력한 다수가 될 것으로 예상된다.

51. 예를 들어, 총 소유 비용 분석에 따르면 대형 클러스터 비용의 약 40%가 H100 GPU 자체이며, 13%는 인피니밴드 네트워킹을 위해 엔비디아에 지불된다. 다만, 자본 비용을 제외하면 GPU가 비용의 약 50%를 차지하고, 네트워킹 비용까지 포함하면 엔비디아가 클러스터 비용의 60% 이상을 차지한다.

52. 그리고 최근 분기에서 마이크로소프트가 전년 대비 자본 지출을 79% 증가시켰음에도 불구하고, AI 클라우드 수요가 여전히 공급을 초과하고 있다!



그림 28:분기별 Nvidia 데이터센터 매출. Thomas Woodside 작성

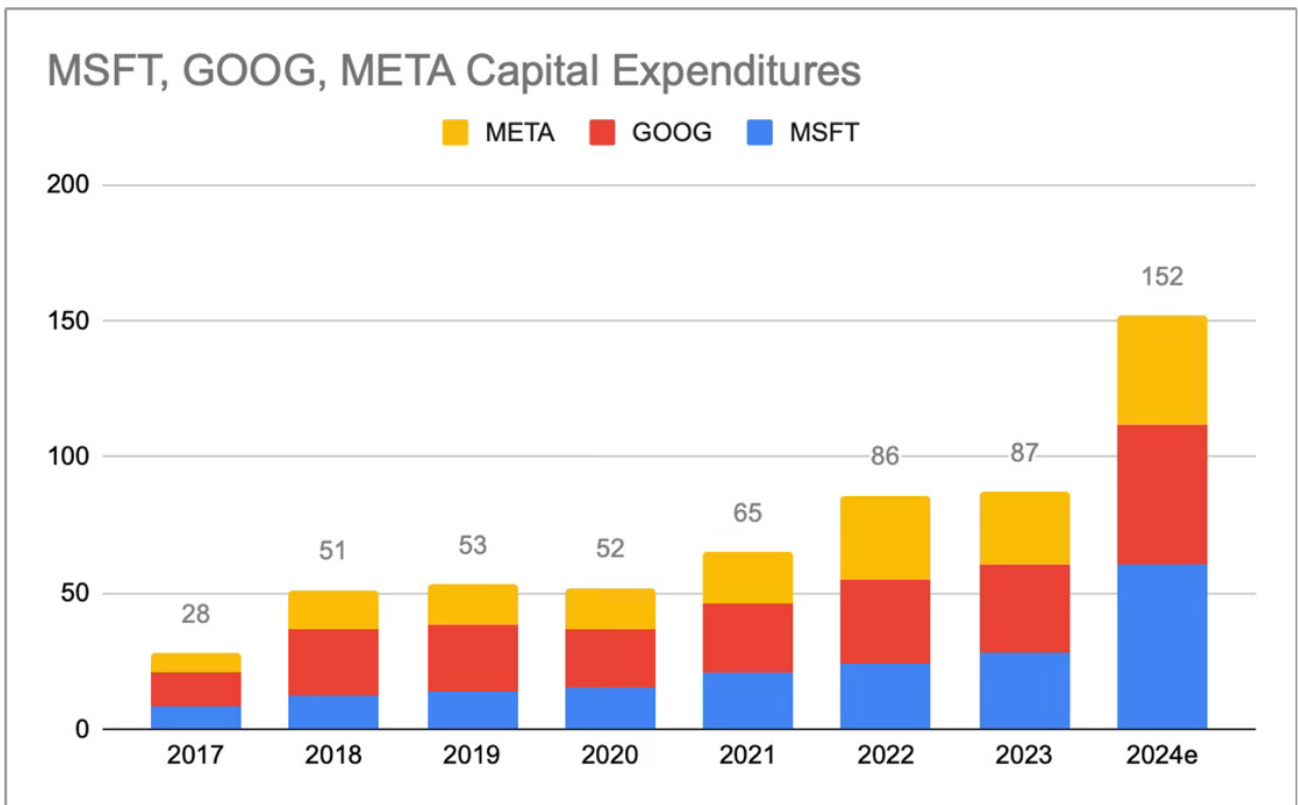
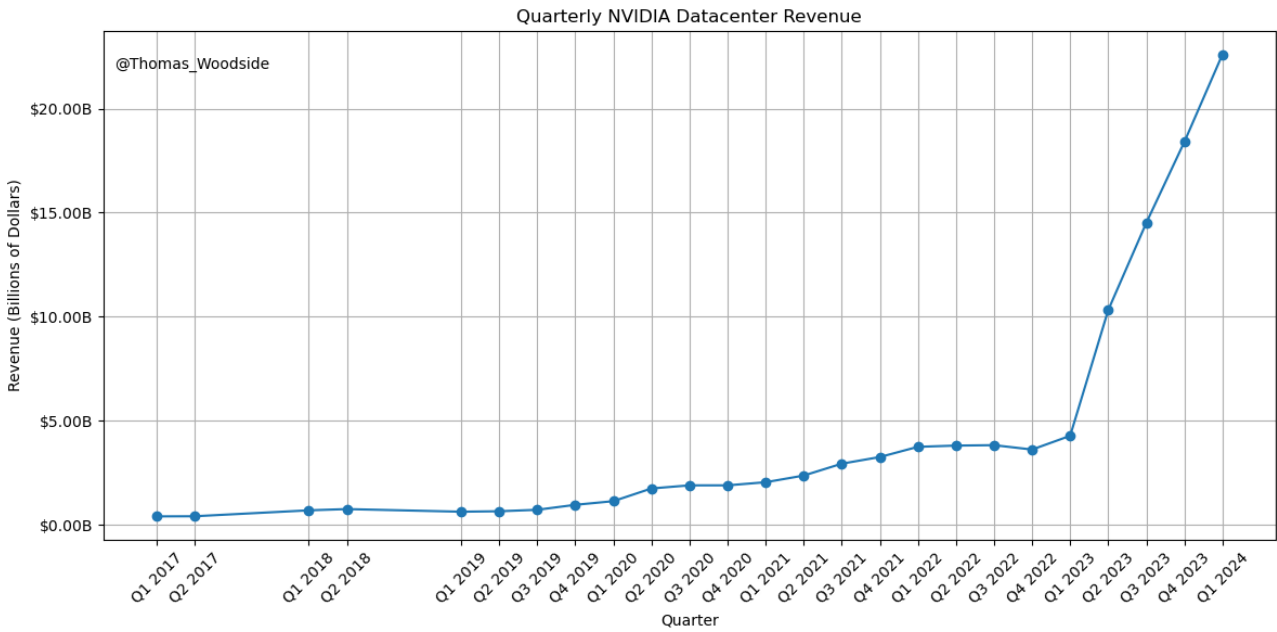


그림 29: ChatGPT가 AI 붐을 촉발한 이후 대형 기술 기업의 자본 지출이 매우 빠르게 증가하고 있음. 출처.

앞으로를 예측해 보자. 내 최선의 추측은 전체 컴퓨팅 투자가 최대 규모의 훈련 클러스터 3배/년 성장보다는 느리게, 예를 들어 2배/년 성장할 것이라는 점이다.⁵³

53. 전 세계 GPU 생산의 더 큰 비율이 미래에는 오늘날보다 최대 규모의 훈련 클러스터에 할당될 가능성이 크다. 예를 들어, 여러 회사가 최첨단 모델 규모 클러스터를 보유하기보다는 소수의 선도 연구소로 통합되는 경향 때문이다. 그리고 이것들은 단지 나만의 특이한 수치가 아니다. AMD는 2027년까지 4,000억 달러 규모의 AI 가속기 시장을 예측했으며, 이는 총 AI 지출이 7,000억 달러 이상임을 의미하는데, 내 수치와 거의 일치한다(그리고 그들은 분명히 나보다 훨씬 덜 '범용 인공지능(AGI) 중심적'이다). 샘 알트만은 다음과 같이 말한다. 전력에 관해서는: 물론 이 모든 것이 미국 내에만 국한되지는 않지만, 참고용 분류를 제공하기 위해서다.

연도 연간 투자액 AI 가속기 출하량 미국 전력 생산량 대비 현재 최첨단 TSMC (달러) (H100s-동등 단위) 전력 비율(%) 웨이퍼 생산량 대비 칩 비율(%) 2024 약 1,500억 달러 약 500만~1,000만대 1~2% 5~10% ~2026 약 5,000억 달러 수천만대 5% 약 25% ~2028 약 2조 달러 1억대 20% 약 100% ~2030 약 8조 달러 수억대 100% 현재 용량의 4배

표 5: 전 세계 AI 총 투자 추세를 앞으로 예상한 대략적인 계산. 계산에 대한 자세한 내용은 부록 참조.

“최대 7조 달러” 규모의 AI 컴퓨팅 용량 구축을 위한 자금 조달을 논의 중이라는 보도가 있었다(이 수치는 널리 조롱받았으나, 여기서 수치를 계산해 보면 덜 터무니없어 보인다...). 어쨌든 이 대규모 확장은 진행 중이다.

과연 이 투자가 이루어질까? 가능할까?

여기서 가정한 투자 규모는 환상적으로 보일 수 있다. 그러나 수요 측과 공급 측 모두 위 경로를 뒷받침할 수 있을 것으로 보인다. 경제적 수익이 투자를 정당화하며, 지출 규모는 새로운 범용 기술에 대해 전례 없는 수준이 아니고, 전력과 칩에 대한 산업 동원도 가능하다.

AI 수익

기업들은 경제적 수익이 이를 정당화할 것으로 기대하면 대규모 AI 투자를 할 것이다.

보고서에 따르면 OpenAI는 2023년 8월에 연간 10억 달러 수익 규모였고, 2024년 2월에는 연간 20억 달러 수익 규모였다. 이는 대략 6개월마다 두 배씩 성장한 것이다. 이 추세가 유지된다면, 차세대 모델에서의 대규모 급증을 반영하지 않더라도 2024년 말~2025년 초에는 약 100억 달러 연간 수익 규모에 도달할 것이다. 한 추정에 따르면 마이크로소프트는 이미 약 50억 달러의 추가 AI 수익을 올리고 있다.

지금까지 AI 투자에서 10배 규모 확장은 필요한 수익을 내는 것으로 보인다. GPT-3.5는 ChatGPT 열풍을 촉발했다. GPT-4 클러스터의 추정 비용 5억 달러는 마이크로소프트와 오픈AI의 연간 수십억 달러 매출로 이미 상쇄되었을 것이다(위 계산 참조). 그리고 수십억 달러 규모의 “2024년형” 훈련 클러스터도 마이크로소프트/오픈AI AI 매출이 100억 달러 이상 매출 속도를 유지한다면 쉽게 회수될 것이다. 이 호황은 투자 주도형이다: 대규모 GPU 주문부터 클러스터 구축, 모델 개발, 배포까지 시간이 걸리며, 오늘 계획 중인 클러스터들은 수년 후에 완성된다. 하지만 마지막 GPU 주문의 수익이 계속 실현된다면 투자는 폭발적으로 증가할 것이고(매출을 앞지를 것이며), 다음 10배 확장도 계속 수익을 낼 것이라는 베팅으로 더 많은 자본이 투입될 것이다.

내가 생각하는 AI 매출의 중요한 이정표는: 대형 기술 기업(구글, 마이크로소프트, 메타 등)이 AI(제품 및 API)로 1000억 달러 매출 속도를 언제 달성할 것인가이다. 이들 기업은 현재 약 1000억~3000억 달러 매출을 올리고 있으므로, 1000억 달러는 그들의 사업에서 매우 큰 비중을 차지하기 시작할 것이다. 매우 단순히 6개월마다 매출이 두 배가 된다고 가정하고, 2025년 초에 100억 달러 매출 속도에 도달한다고 하면, 이 이정표는 2026년 중반에 달성될 것으로 보인다.

이것이 다소 무리처럼 보일 수 있지만, 나는 이 이정표에 도달하는 데 놀랄 만큼 적은 상상력이 필요하다고 생각한다. 예를 들어, 마이크로소프트 오피스 유료 구독자가 약 3억 5천만 명인데, 이 중 3분의 1이 AI 추가 기능에 월 100달러를 지불할 의향이 있다고 할 수 있을까? 평균 근로자에게는 한 달에 몇 시간의 생산성 향상에 불과하다; 앞으로 몇 년 내에 이를 정당화할 만큼 강력한 모델을 만드는 것은 매우 가능해 보인다.⁵⁴

이로 인한 파급 효과는 과소평가하기 어렵다. 이는 AI 제품이 미국 최대 기업들의 가장 큰 매출 동력이자 단연 최대 성장 분야가 됨을 의미한다. 이들 기업의 전체 매출 성장 전망은 급등할 것이다. 주식 시장도 이에 따라 움직일 것이며, 곧 최초의 10조 달러 기업이 등장할 수도 있다. 이 시점에서 대형 기술 기업들은 AI 확장에 수천억 달러(최소한 수백억 달러)를 투자할 준비가 되어 있을 것이다. 우리는 아마도 최초의

수백억 달러 규모의 기업 채권 발행 당시.⁵⁵

역사적 선례

2027년까지 연간 1조 달러 규모의 AI 투자는 터무니없어 보인다. 하지만 다른 역사적 참조 사례를 살펴볼 가치가 있다:

- 맨해튼 프로젝트와 아폴로 프로젝트는 최고 자금 지원 시기에 GDP의 0.4%, 즉 오늘날 약 1000억 달러를 연간 투자했다(놀랍게도 적은 금액!). AI 투자가 연간 1조 달러라면 GDP의 약 3%에 해당한다.
- 1996년부터 2001년까지 통신사들은 오늘날 가치로 거의 1조 달러를 인터넷 인프라 구축에 투자했다.
- 1841년부터 1850년까지 영국 민간 철도 투자는 당시 영국 GDP의 누적 약 40%에 달했다. 미국 GDP의 비슷한 비율은 10년간 약 11조 달러에 해당한다.
- 녹색 전환에 수조 달러가 투입되고 있다.
- 급성장하는 경제는 종종 GDP의 높은 비율을 투자에 쓴다; 예를 들어, 중국은 20년간 GDP의 40% 이상을 투자에 사용해왔다(미국 GDP 기준으로 연간 약 11조 달러에 해당).
- 역사상 가장 긴급한 국가안보 상황인 전시에는 국가 노력을 위한 차입이 GDP의 막대한 비율을 차지했다. 제1차 세계대전 당시,

영국과 프랑스, 독일은 GDP의 100% 이상을 차입했으며, 미국은 20% 이상을 차입했다; 제2차 세계대전 동안 영국과 일본은 GDP의 100% 이상을 차입했고 미국은 GDP의 60% 이상을 차입했다(오늘날 가치로 17조 달러 이상에 해당).

2027년까지 연간 1조 달러 규모의 총 AI 투자는 매우 극적인 규모로, 지금까지의 가장 큰 자본 투자 중 하나가 될 것이지만 전례 없는 일은 아니다. 그리고 10년 말까지 1조 달러 규모의 개별 훈련 클러스터도 가능해 보인다.

⁵⁶

전력 아마도 공급 측면에서 가장 큰 제약은 전력일 것이다. 이미 가까운 미래 규모(2026년 1GW, 특히 2028년 10GW)에서는 전력이 제약 요인이 되었다: 여분의 용량이 거의 없으며, 전력 계약은 보통 장기 고정되어 있다. 예를 들어, 새로운 기가와트급 원자력 발전소를 건설하는 데는 10년이 걸린다. (기술 기업들이 기가와트급 전력 계약을 위해 알루미늄 제련 회사를 인수하는 모습을 언제쯤 보게 될지 궁금하다.⁵⁷)

54. 저에게 큰 불확실성은 기술 확산 및 도입 지연입니다. 중간 단계 AGI 이전 모델들이 기업 업무 프로세스에 제대로 통합되기까지 시간이 걸릴 가능성이 높으며, 역사적으로 새 범용 기술의 생산성 혁신이 완전히 발현되는 데 시간이 걸렸습니다. 앞서 논의한 '소닉 붐'과 연결됩니다: 모델들이 '구속 해제'되어 에이전트나 원격 근무자처럼 작동하기 시작하면 배포가 훨씬 쉬워질 것입니다. GPT 챗봇에서 25% 생산성 향상을 얻으려면 업무 흐름을 완전히 재구성해야 하는 것이 아니라, 새로운 동료를 온보딩하여 함께 일하듯 단순히 엔지니어를 대체하면 됩니다. 심지어 나중에는 공장 설계를 전면 개조할 필요 없이 휴머노이드 로봇만 투입하면 됩니다. 다만, 얼마나 빠르게 '구속 해제'가 이루어지느냐에 따라 경제적 가치와 수익에 불연속성이 발생할 수도 있습니다.

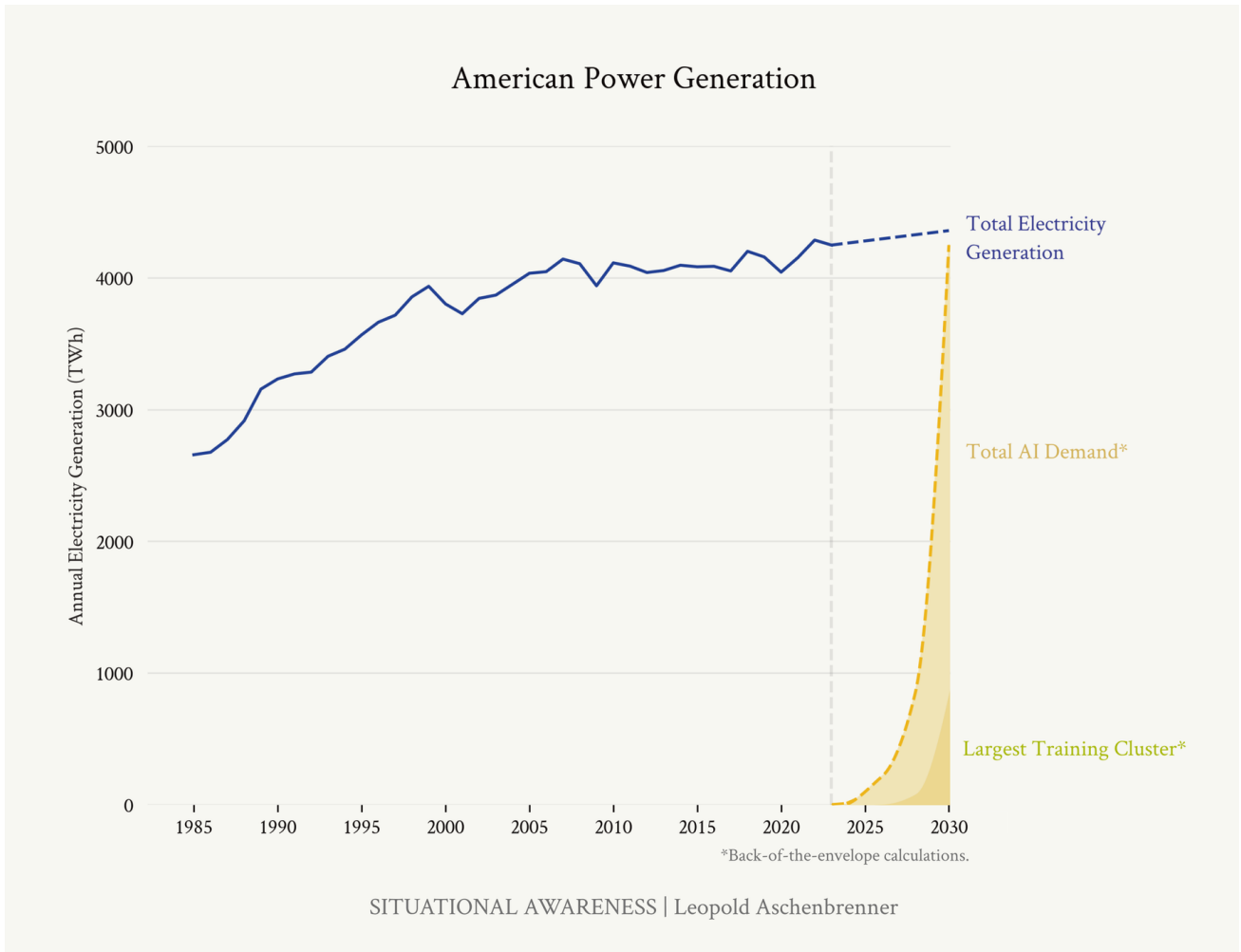


그림 30:미국 전체 전력 생산 추세와 AI 전력 수요에 대한 대략적인 추정치 비교.

미국 전체 전력 생산량은 지난 기간 동안 5%도 채 성장하지 못했다.

지난 10년간, 유틸리티 기업들은 AI에 대해 점점 더 흥미를 느끼기 시작했다(앞으로 5년간 2.6% 성장 예상에서 이제는 4.7%로 추정!). 하지만 그들이 예상하는 수준은 다가올 상황에 비하면 거의 미미하다. 1조 달러 규모, 100GW 클러스터만 해도 6년 내에 현재 미국 전력 생산량의 약 20%를 필요로 할 것이다; 대규모 추론 용량까지 합치면 수요는 훨씬 더 커질 것이다.

대부분 사람들에게 이것은 전혀 불가능해 보인다. 일부는 중동의 독재국가들이 무한한 전력과 거대한 클러스터를 제공하며 그들의 통치자들이 범용 인공지능(AGI) 테이블에 앉을 수 있도록 하려 한다고 베팅하고 있다.

하지만 미국 내에서도 충분히 가능하다: 우리는 풍부한 천연가스를 보유하고 있다.⁵⁸

- 100GW 클러스터조차도 놀랍도록 실현 가능하다.

55. 금리 변동이 어떻게 될지 흥미로운 것이다... 타일러 코언(Tyler Cowen)을 참고하라; 초우(Chow), 마즐리시(Mazlish), 할페린(Halperin)도 참고. \$1000억을 넘어서면 윤곽을 파악하기가 더 어려워진다. 하지만 우리가 진정 범용 인공지능(AGI)으로 가는 길에 있다면, 수익은 분명 있을 것이다. 화이트칼라 근로자들은 전 세계적으로 매년 수십조 달러의 임금을 받는다; 원격 근로자 한 명이 화이트칼라/인지 직업의 일부만 자동화해도 (예를 들어, 진정 자동화된 AI 코더를 상상해 보라) 조 단위 클러스터 비용을 충당할 수 있다. 최소한 국가안보 측면에서 정부 프로젝트를 추진해 국가 자원을 묶어 AGI 경쟁에 나설 수도 있다(후후 설명).

- 현재 마르셀러스/유티카 세일(펜실베이니아 인근)만 해도 하루 약 360억 입방피트의 가스를 생산 중이다; 이는 발전기를 통해 연속적으로 약 150GW를 생산할 수 있는 양이다(복합 사이클 발전소는 효율이 높아 250GW까지 출력 가능).

- 100GW 클러스터를 위해서는 약 1200개의 신규 유정이 필요하다.⁵⁹ 각 시추기는 한 달에 약 3개의 유정을 시추할 수 있으므로, 40대의 시추기(현재 마르셀러스 시추기 수)로 1년 이내에 100GW 생산 기반을 구축할 수 있다.⁶¹⁶⁰

- 더 일반적으로, 미국 천연가스 생산량은 10년간 두 배 이상 증가했다; 이 추세를 계속 유지하기만 해도 수조 달러 규모의 데이터센터 여러 곳에 전력을 공급할 수 있다.⁶²

- 더 어려운 부분은 충분한 발전기/터빈을 건설하는 것인데; 이는 간단하지 않지만 실현 가능해 보인다.

약 1000억 달러의 자본 지출(capex)⁶³로 100GW의 천연가스 발전소

우리는 AGI 데이터센터를 잔인하고 변덕스러운 독재자들의 손아귀에 있는 나도 청정 에너지를 선호하지만—중등으로 몰고 갈 것임. 하지만 이것은 미국 국가안보에 너무 중요함. 이 문제를 해결하려면 새로운 수준의 결단력이 필요함. 전력 문제는 반드시, 그리고 반드시 해결될 수 있고 해결되어야 함.

칩

사람들이 AI 공급 제약을 생각할 때 보통 칩을 떠올리지만, 전력보다 65 이는 칩 제조가 심각한 제약에 작은 제약일 가능성이 큼. AI 칩의 전 세계 생산량은 TSMC 최첨단 생산의 직면하기 전에 전력 요구량이 왜 이렇게 많이 증가하는지에 대한 아직도 매우 작은 비율로, 10% 미만일 가능성이 큼. AI가 TSMC 생산에서 흥미로운 질문을 제기함. 더 큰 비중을 차지함에 따라 성장 여지가 많음. 간단한 답은 데이터센터는 거의 최대 전력으로 계속 가동되는 반면, 현재 생산되는 대부분 칩은 많이 실제로 2024년 AI 칩 생산량(~500만~1000만 H100 상당)은 이미 수천억 달러 유휴 상태임. 현재 스마트폰이 주요 칩 수요의 거의 절반이지만, 규모 클러스터(모두 한 클러스터에 투입된다면)에 거의 충분함. 순수 논리 웨이퍼 면적당 에너지 사용량은 훨씬 적음(트랜지스터를 직렬 연산과 파운드리 관점에서 보면 TSMC 연간 생산량의 거의 100%가 이미 수조 달러 에너지 효율성으로 교환)이며, 스마트폰은 대부분 유휴 상태여서 규모 클러스터를 지원할 수 있음(모두 한 데이터센터에 투입된다면).⁶⁵ 물론, 활용률이 낮음. AI 혁명은 트랜지스터를 훨씬 더 열심히 작동시키고, 모두 TSMC 전체가 AI에 투입될 수 없고, 1년간 AI 칩 생산량이 한 훈련 클러스터에 끊임없이 가동되는 고성능 AI 데이터센터에 전념하게 만듦. 배터리 구동/에너지 모두 투입되지도 않을 것임. AI 칩 총 수요(HT Carl Shulman 지적 포함) 절약 장치 대신.

추론과 다중 플레이어를 포함한 AI용 컴퓨팅 수요는 2030년까지 TSMC의 현재 최첨단 논리 칩 총 생산 능력의 수배에 이를 것이다. TSMC는 지난 5년간 거의 두 배로 성장했으며⁶⁶, AI 칩 수요를 충족하려면 확장 속도를 최소 두 배 이상으로 높여야 할 것이다. 대규모 신규 파운드리 투자가 필수적이다.

원시 논리 파운드리가 제약이 되지 않더라도, 칩-온-웨이퍼-온-기판(CoWoS) 첨단 패키징(칩을 메모리와 연결하는 기술로 TSMC, 인텔 등에서 생산)과 HBM 메모리(수요가 엄청남)는 현재 AI GPU 확장에 이미 주요 병목이

56. 그리고 더 먼 미래에는, 만약 범용 인공지능(AGI)이 경제 성장에 실질적인 증가를 가져온다면, 연간 10조 달러 이상의 투자가 현실적으로 가능해질 것이다—참고 대상은 고성장 시기의 국가 투자율이다.

57. “2011년 이후, 알루엣 제련소는 최대 생산 용량에서 930MW의 전력을 사용하고 있다.”

다. 이들은 순수 논리 칩과 달리 AI에 더 특화되어 있어 기존 생산 능력이 적다. 단기적으로는 이들이 더 많은 GPU를 생산하는 데 주된 제약이 될 것이며, AI가 확장됨에 따라 큰 제약으로 작용할 것이다. 그럼에도 불구하고 이들은 비교적 “확장하기 쉬운” 분야이다. 올해 TSMC가 CoWoS 생산을 대규모로 확장하기 위해 완전히 새로운 시설(그린필드 파운드리)을 건설하는 모습을 보는 것은 놀라웠다(엔비디아는 부족 현상을 우회하기 위해 CoWoS 대체 기술도 모색 중이다).

신규 TSMC 기가파브(기술적 경이)는 약 200억 달러의 자본 지출로 월 10만 웨이퍼 시작량을 생산한다. 10년 말까지 수억 대의 AI GPU를 생산하려면 TSMC는 이와 같은 시설을 수십 개 건설해야 하며, 메모리, 첨단 패키징, 네트워킹 등에도 대규모 투자가 필요하다. 이들은 자본 지출의 큰 부분을 차지할 것이다. 총합은 1조 달러 이상의 자본 지출에 이를 수 있다. 매우 강도 높은 작업이지만 가능하다. (가장 큰 장애물은 실현 가능성이 아니라 TSMC가 시도조차 하지 않는 것일 수 있다—TSMC는 아직 AI 확장에 대해 적극적이지 않은 것으로 보인다! 그들은 AI가 “겨우” 연평균 50% 성장할 것이라고 생각한다.)

CHIPS 법안과 같은 최근 미국 정부의 노력은 AI 칩 생산을 미국 내로 이전하려는 시도이다(대만 위기 대비 보험 차원). AI 칩 생산을 미국 내로 이전하는 것은 바람직하지만, 실제 AGI가 존재하는 데이터센터를 미국에 두는 것만큼 중요하지는 않다. 해외에 칩 생산 시설이 있는 것은 해외에 우라늄 광산이 있는 것과 같다면, AGI 데이터센터가 해외에 있는 것은 실제 핵무기가 해외에서 제조되고 저장되는 것과 같다. 기능 장애를 고려할 때

미국 내에서 파운드리(fab)를 구축하는 데 실제로 드는 비용과 경험을 고려할 때, 저는 미국 내 데이터센터를 우선시하면서 파운드리 프로젝트는 일본과 한국 같은 민주주의 동맹국에 더 무게를 두어야 한다고 생각합니다. 그곳에서의 파운드리 구축은 훨씬 더 실용적인 것으로 보입니다.

민주주의 클러스터

10년이 끝나기 전에 수조 달러 규모의 컴퓨터 클러스터가 구축될 것입니다. 유일한 질문은 그것들이 미국에서 구축될 것인가 하는 점입니다. 일부는 특히 중동에서 구축하는 데 베틱하고 있다고 소문이 돌고 있습니다. 우리는 정말로 맨해튼 프로젝트의 인프라가 번덕스러운 중동 독재 정권에 의해 통제되는 것을 원합니까?

오늘날 계획 중인 클러스터들은 단순한 ‘멋진 대형 기술 제품 클러스터’가 아니라 AGI와 초지능이 훈련되고 운영될 클러스터일 가능성이 큼니다. 국가 이익은 이들이 미국(또는 가까운 민주주의 동맹국)에서 구축되도록 요구합니다. 그렇지 않으면 돌이킬 수 없는 안보 위험이 발생합니다. AGI 모델 가중치가 도난당할 위험⁶⁷(그리고 아마도 중국으로 반출될 위험)이 있으며, AGI 경쟁이 치열해질 때 이 독재 정권들이 데이터센터를 물리적으로 점령하여 AGI를 직접 구축하고 운영할 위험도 있습니다. 이러한 위험이 암묵적으로만 행사되더라도 AGI와 초지능이 불쾌한 독재자의 번덕에 휘둘리게 됩니다. 미국은 70년대 중동에 대한 에너지 의존을 뼈저리게 후회했고, 그들의 지배에서 벗어나기 위해 매우 노력했습니다. 우리는 같은 실수를 반복할 수 없습니다.

58. 이 추정치를 도와준 오스틴 버논(Austin Vernon, 개인 서신)에 감사드립니다. • 10GW 클러스터에 전력을 공급하는 데는 미국 천연가스 생산량의 몇 퍼센트만 필요하며 빠르게 진행할 수 있다.

59. 신규 유정은 하루 약 0.01 BCF 생산합니다.

61. 다만 10개월보다 긴 기간 동안 굴착기 수를 적게 늘리고 점진적으로 확장하는 쪽이 더 효율적일 것입니다.

60. 각 유정은 수명 동안 약 20 BCF를 생산하며, 매달 두 개의 신규 유정이 고갈된 매장량을 대체합니다. 즉, 생산 유지에 필요한 굴착기는 한 대뿐입니다.

클러스터는 미국 내에서 구축할 수 있으며, 반드시 미국에서 구축되도록 우리의 역량을 모아야 합니다. 미국의 국가 안보가 중동에서 흘러들어오는 자금, 복잡한 규제, 심지어는 존경할 만한 기후 약속보다 우선해야 합니다. 우리는 진정한 체제 경쟁에 직면해 있습니다—필요한 산업 동원이 ‘상명하달’식 독재 체제에서만 가능한가? 미국 기업이 제약에서 벗어난다면, 미국은 그 어느 곳보다도 잘 구축할 수 있습니다(적어도 보수 성향 주에서는). 천연가스를 기꺼이 사용하거나 최소한 광범위한 규제 완화 정책—NEPA 면제, FERC 개혁 및

연방 차원의 전송 허가, 공익사업 규제 무시, 연방 권한을 사용해 토지와 통행권 확보—이는 국가안보 우선순위다.

어쨌든: 지수적 성장은 이제 본격적으로 진행 중이다.

“예전”에는 AGI가 아직 금기어였을 때, 몇몇 동료들과 나는 AGI로 가는 경로가 어떻게 될지 이론적 경제 모델을 만들곤 했다. 이 모델들의 한 특징은 가상의 “AI 각성” 순간이었다. 이때 세계가 이 모델들이 얼마나 강력할 수 있는지 깨닫고 투자를 급격히 늘리기 시작하며, 최대 규모의 훈련에 GDP의 여러 %를 쏟아붓는 시점이었다.

그때는 먼 미래처럼 보였지만, 그 시간이 왔다. 2023년이 바로 “AI 각성”이었다.⁶⁸ 무대 뒤에서는 가장 놀라운 기술 자본 가속이 시작되었다.

2023년 3월, 나는 분명히 내 화이트보드에 “THE TAKEOFF HAS STARTED(이륙이 시작되었다)”라고 썼다.

G-포스에 대비하라.

(이 모든 것이 NVDA/TSM 등 주가에 어떤 의미인지 독자 여러분이 직접 생각해 보시길 바란다. 힌트: 상황 인식이 있는 이들은 훨씬 낮은 가격에 매수했지만, 아직도 완전히 반영된 것은 아니다.)⁶⁹

IIIb. 연구소 봉쇄: 범용 인공지능(AGI) 보안

국내 주요 AI 연구소들은 보안을 사후 고려 사항으로 취급하고 있다. 현재 이들은 사실상 AGI의 핵심 비밀을 중국공산당(CCP)에게 은쟁반에 올려주는 꼴이다. 국가 행위자 위협으로부터 AGI 비밀과 모델 가중치를 보호하는 일은 엄청난 노력이 필요하며, 우리는 그 궤도에 있지 않다.

그들은 저녁에 와이그너 사무실에서 만났다. 월러는 “실라드가 콜롬비아 데이터를 개괄했고, 그로부터 최소 두 개의 2차 중성자가 중성자 유도 핵분열에서 나온다는 예비 징후를 보고했다. 이것이 핵폭발이 확실히 가능하다는 뜻이 아니냐?”고 전한다. 보어는 반드시 그렇지 않다고 반박했다.

텔러는 “우리는 그를 설득하려 했다. 핵분열 연구를 계속하되 결과를 발표하지 말자고. 나치가 이를 알게 되어 먼저 핵폭발을 일으키는 일을 막기 위해 결과를 비밀로 해야 한다고.” 썼다.

62. 천연가스 1입방피트는 약 0.13 kWh를 생산한다. 2020년 미국 세일 가스 생산량은 하루 약 700억 입방피트였다. 생산량이 다시 두 배가 되고, 추가 용량이 모두 컴퓨팅 클러스터에 할당된다 가정하면 연간 3322 TWh의 전력이 생산된다. 이는 거의 4개의 100GW 클러스터에 충분한 전력이다.

63. 천연가스 발전소의 투자비용은 kW당 1000달러 이하로 보이며, 100GW 용량의 투자 비용은 약 1000억 달러가 됩니다.

65. 칩 생산이 본격적인 제약을 받기 전에 전력 수요가 급격히 증가하는 흥미로운 질문입니다. 간단한 대답은 데이터센터는 최대 전력에 가깝게 연속 가동하는 반면, 현재 생산되는 대부분 칩은 상당 시간 휴지 상태라는 점입니다. 현재 스마트폰이 주 반도체 수요의 절반 가량을 차지하지만, 공정 밀도 대비 효율이 높고 스마트폰은 주로 휴지 상태라 에너지 사용량이 적습니다. AI 혁명은 트랜지스터를 훨씬 더 세계 작동하게 하여, 배터리 전력 절약용 저성능 장치 대신 고성능 AI 데이터센터에 전적으로 투입하는 것입니다. Carl Shulman에게 이 점 감사드립니다.

“보어는 우리가 핵에너지를 생산하는 데 결코 성공하지 못할 것이며, 물리학에 비밀주의가 도입되어서는 안 된다고 주장했다.”

『원자폭탄의 탄생』, 430쪽

현재의 방향대로라면, 중국의 주요 AGI 연구소들은 베이징이나 상하이보다 아니라 샌프란시스코와 런던에 위치할 것이다. 몇 년 내에 AGI 비밀이 미국의 가장 중요한 국가 방위 비밀임이 명확해질 것이다. 이는 B-21 폭격기나 콜럼비아급 잠수함 설계도와 동등한 대우를 받아야 하며, 말할 것도 없이...

“핵 비밀”처럼—하지만 오늘날 우리는 그것들을 무작위 SaaS 소프트웨어 다루듯이 취급하고 있다. 이 속도라면, 우리는 사실상 초지능을 중국공산당(CCP)에게 그냥 넘겨주고 있는 셈이다.

우리가 투자할 수조 달러, 미국 산업 역량의 동원, 가장 뛰어난 인재들의 노력—중국이나 다른 나라가 모델 가중치(완성된 AI 모델, 범용 인공지능(AGI)도 결국 컴퓨터상의 큰 파일에 불과하다)나 핵심 알고리즘 비밀(AGI를 구축하는 데 필요한 핵심 기술적 돌파구)을 단순히 훔칠 수 있다면, 그 모든 것은 무의미하다.

미국의 주요 AI 연구소들은 스스로 AGI를 구축하고 있다고 선언한다: 그들은 자신들이 개발하는 기술이 10년 내에 미국이 만든 가장 강력한 무기가 될 것이라고 믿는다. 하지만 그들은 그것을 그렇게 취급하지 않는다. 그들은 보안 노력을 “무작위 기술 스타트업”과 비교할 뿐, “핵심 국가 방위 프로젝트”와 비교하지 않는다. AGI 경쟁이 치열해지고—초지능이 국제 군사 경쟁에서 결정적인 역할을 할 것이 분명해지면서—우리는 외국 스파이 활동의 전면적인 위협에 직면해야 할 것이다. 현재 연구소들은 스크립트키디(script kiddies) 정도의 공격도 제대로 방어하지 못하며, “북한 수준의 보안”은 커녕 중국 국가안전부가 전력을 다해 공격해오는 상황에 대비할 준비도 되어 있지 않다.

그리고 이것은 몇 년 후의 문제가 아니다. 물론, GPT-4 가중치가 도난당해도 누가 신경 쓰겠는가—가중치 보안에서 진짜 중요한 것은 AGI 가중치를 안전하게 지키는 것이므로, 몇 년의 시간이 있다고 말할 수도 있다. (하지만 만약 우리가 2027년에 AGI를 구축한다면, 정말 서둘러야 한다!) 그러나 AI 연구소들은 지금 바로 알고리즘 비밀—즉 AGI를 위한 핵심 기술적 돌파구, 일종의 청사진(특히 LLM을 넘어 데이터 장벽을 극복하기 위한 RL/자기 대결/합성 데이터 등 “다음 패러다임”)을 개발하고 있다. 알고리즘 비밀에 대한 AGI 수준의 보안은 가중치에 대한 AGI 수준의 보안보다 수년 앞서 필요하다. 이러한 알고리즘 돌파구는 몇 년 후 10배 또는 100배 더 큰 클러스터보다 더 중요하다—이는 미국 정부가 (선견지명 있게!) 강력히 추진해온 컴퓨트 수출 통제보다 훨씬 더 중대한 문제다. 지금 당장, 이러한 비밀을 훔치기 위해 극적인 스파이 작전을 벌일 필요도 없다: 그냥 샌프란시스코 파티에 가거나 사무실 창문을 들여다보면 된다.

66. (수익을 군사 지표로 사용함.)

67. 물리적 접근권이 있으면 사이드채널 공격으로 가중치를 탈취하기가 훨씬 쉽습니다!

68. 저는 2023년 3월에 “이륙이 시작되었다(THE TAKEOFF HAS STARTED)”라고 화이트보드에 썼던 기억이 선명합니다.

69. 주류 매도측 애널리스트들은 2024년부터 2025년까지 엔비디아 매출이 연간 10-20% 성장할 것으로 보고, 2025년 매출을 약 1,200억 ~1,300억 달러로 예상하는 경향이 있다(최근까지도 그랬다). 미친 전망이다! 엔비디아가 2025년에 2,000억 달러 이상의 매출을 올릴 것이라는 점은 이미 꽤 명백하다.

오늘 우리의 실패는 곧 돌이킬 수 없게 될 것이다: 향후 12~24개월 내에 우리는 주요 범용 인공지능(AGI) 돌파구를 중국공산당(CCP)에게 유출할 것이다. 이것은 10년이 끝나기 전에 국가안보 체제의 가장 큰 후회가 될 것이다.

권위주의 국가들에 맞서 자유 진영을 지키는 것이 걸려 있으며, 건강한 우위는 AI 안전을 제대로 확보할 수 있는 여유를 제공하는 필수 완충장치가 될 것이다. 미국은 AGI 경쟁에서 우위를 점하고 있다. 그러나 우리가 곧바로 보안 문제를 심각하게 다루지 않으면 이 우위를 포기하게 될 것이다. 지금 이 문제에 착수하는 것이 아마도 AGI가 잘 작동하도록 보장하기 위해 오늘 우리가 해야 할 가장 중요한 일일 것이다.

국가 행위자를 과소평가하지 마라

너무 많은 똑똑한 사람들이 첩보 활동을 과소평가한다.

국가와 그 정보기관의 역량은 매우 강력하다. 정상적인, 전면적인 AGI 경쟁이 아닌 시기에도 (그리고 우리가 공개적으로 아는 바는 적지만), 국가(또는 덜 발전된 행위자)들은 다음과 같은 일을 해낼 수 있었다:

- 단지 전화번호만으로 원하는 iPhone과 Mac을 제로 클릭 해킹,
- 에어갭이 적용된 핵무기 프로그램 침투,
- 구글 소스 코드 수정,
- 연평균 7년이 걸려 발견되는 수십 건의 제로데이 취약점 발견,
- 주요 기술 기업을 대상으로 스피어피싱,
- 직원 기기에 키로거 설치,
- 암호화 체계에 트랩도어 삽입,
- 전자기 방출이나 진동을 통한 정보 탈취,
- 컴퓨터에서 발생하는 잡음만으로 비디오 게임 지도에서 자신의 위치를 파악하거나 비밀번호를 훔치기,
- 원자력 발전소와 같은 민감한 시스템에 직접 접근하기,
- 미국 정부(USG)로부터 2,200만 개의 보안 승인 파일을 빼내기,
- HVAC 시스템에 취약점을 심어 1억 1,000만 고객의 금융 정보를 노출시키기,
- 대규모로 컴퓨터 하드웨어 공급망을 침해하기,
- 주요 기술 기업과 미국 정부가 사용하는 소프트웨어 의존성 업데이트에 악성 코드를 몰래 삽입하기,
- … 스파이를 심거나 직원을 유혹, 회유, 협박하는 것은 말할 것도 없고(이는 대규모로 효과적으로 이루어지지 않더라도),
- … 특수부대 작전 등과 같은 일(상황이 정말 심각해질 때)은 더 말할 것도 없다.

정보 기관과 맞서 싸울 때 우리가 직면한 상황을 더 잘 이해하기 위해, 소련 GRU(군사 정보) 탈북자의 책 《Inside the Aquarium》을 강력히 추천한다.⁷⁰

이미 중국은 광범위한 산업 스파이 활동을 벌이고 있으며, FBI 국장은 중국공산당(CCP)이 “모든 주요 국가를 합친 것보다 더 큰” 해킹 작전을 수행하고 있다고 밝혔다. 그리고 불과 몇 달 전, 법무장관은 중국 국적자가 구글에서 핵심 AI 코드를 훔쳐 중국공산당에 가져가려다 체포되었다고 발표했다(2022/23년경이며, 아마도 빙산의 일각에 불과하다).⁷¹

하지만 이것은 시작에 불과하다. 우리는 적들이 향후 몇 년 내에 범용 인공지능(AGI)에 “눈을 뜰” 것에 대비해야 한다. AI는 전 세계 모든 정보 기관의 최우선 순위가 될 것이다.

그 상황에서 그들은 AI 연구소에 침투하기 위해 비상한 수단을 동원하고 어떤 대가도 치를 준비가 되어 있을 것이다.

“딩(DING)은 구글 소스 파일에서 데이터를 복사해 구글에서 지급한 맥북 노트북의 애플 노트 애플리케이션에 붙여넣어 이 파일들을 빼냈다. 딩은 애플 노트를 PDF 파일로 변환한 후 구글 네트워크에서 딩 계정 1로 업로드했다. 이 방법 덕분에 딩은 즉각적인 탐지를 피할 수 있었다.” (기소장 발췌)

그가 잡힌 이유는 중국에서 유명 스타트업에 바로 시작하는 등 다른 어리석은 행동을 많이 했기 때문이다. 이로 인해 사람들이 의심했고(나중에는 미국으로 돌아오기도 했다).

위협 모델

우리가 보호해야 할 두 가지 핵심 자산이 있다: 모델 가중치(특히 범용 인공지능(AGI)에 가까워질수록 중요하지만, 이를 제대로 준비하고 실현하는 데는 수년이 걸린다)와 알고리즘 비밀(어제부터 시작된).

모델 가중치

AI 모델은 서버에 저장된 거대한 숫자 파일에 불과하다. 이 파일은 도난당할 수 있다. 적이 당신의 수조 달러, 가장 똑똑한 인재들, 수십 년간의 노력을 맛먹으려면 이 파일을 훔치기만 하면 된다. (만약 나치가 로스앨러모스에서 만든 모든 원자폭탄의 정확한 복제품을 얻었다고 상상해 보라.)

모델 가중치를 안전하게 지키지 못하면, 우리는 단지 중국공산당(CCP)을 위해 AGI를 만드는 셈이다(현재 AI 연구소 보안 수준을 고려하면 북한도 포함된다).

국가 간 경쟁을 떠나서도, 모델 가중치 보안은 AI 재앙을 막는 데 필수적이다. 나쁜 행위자(예: 테러리스트나 불량 국가)가 모델을 훔쳐 안전 장치를 우회해 원하는 대로 사용할 수 있다면, 우리의 모든 걱정과 보호 조치는 무의미해진다. 초지능이 발명할 수 있는 새로운 대량살상무기는 수십 개의 불량 국가로 빠르게 확산될 것이다. 게다가 보안은 통제 불능 또는 정렬되지 않은 AI 시스템에 대한 첫 번째 방어선이기도 하다(만약 우리가 민감구획

70. 한 가지 예를 들자면: ~스파이 아카데미를 졸업하고 해외에 파견되기 전에, 예비 스파이들은 국내에서 자신의 능력을 증명해야 했다. 그들은 소련 과학자로부터 비밀 정보를 획득해야 했으며, 국가 비밀을 누설할 경우 처벌은 당연히 사형이었다. 즉, ~스파이 아카데미를 졸업한다는 것은 동료 국민 한 명을 사형에 처하도록 선택하는 것을 의미했다. 책 추천에 대해 일리아 수츠케버(Ilya Sutskever)에게 감사한다.

71. 참고로, 기소장에는 구글에서조차 보안 회피가 얼마나 쉬운지 잘 보여주는 사례가 있다. 구글은 아마도 AI 연구소 중 가장 뛰어난 보안을 갖추고 있을 텐데(구글의 수십 년간 보안 인프라 투자 덕분에). 코드를 훔치기 위해 탐지 없이 한 일은 애플 노트에 코드를 붙여넣고 PDF로 내 보내는 것이 전부였다!

정보시설(SCIF) 같은 에어갭 클러스터에서 먼저 구축하고 테스트하지 않아 불량 초지능을 통제하지 못했다면 얼마나 어리석게 느껴질까?).

현재는 모델 가중치 보안이 크게 중요하지 않다. GPT-4를 기반 레시피 없이 훔친다고 해도 CCP에 큰 도움이 되지 않는다. 하지만 AGI가 등장하는 몇 년 후에는 정말 중요해질 것이다. 진정으로 엄청난 힘을 가진 시스템이 등장할 때 말이다.

내가 가장 밤잠을 설치게 하는 시나리오 중 하나는 중국이나 다른 적대 세력이 지능 폭발 직전의 자동화 AI 연구자 모델 가중치를 훔치는 경우다. 중국은 이를 즉시 사용해 AI 연구를 자동화할 수 있다.

(설령 이전에 훨씬 뒤처져 있었다라도)—그리고 그들만의 지능 폭발을 시작할 것이다. 그것만으로도 AI 연구를 자동화하고 초지능을 구축하는 데 충분하다. 미국이 가진 어떤 우위도 사라질 것이다.

더욱이, 이것은 즉시 우리를 존재론적 경쟁에 몰아넣을 것이다; 초지능의 안전을 보장할 여유가 사라진다. 중국 공산당(CCP)은 아마도 가능한 한 빠르게 지능 폭발을 서두르려 할 것이다—초지능에서 몇 달의 우위만으로도 결정적인 군사적 이점을 의미할 수 있기 때문에—그 과정에서 어떤 책임 있는 미국 범용 인공지능(AGI) 노력이 취하려는 모든 안전 조치를 건너뛸 수 있다. 우리는 또한 완전한 CCP 지배를 피하기 위해 지능 폭발을 서둘러야 할 것이다. 설령 미국이 결국 간신히 앞서 나간다고 해도, 여유가 사라진다는 것은 AI 안전에 엄청난 위험을 감수해야 함을 의미한다.

오늘날 모델 가중치를 보호할 충분한 보안 수준과는 거리가 멀다. 구글 딥마인드(아마도 구글 인프라 덕분에 가장 뛰어난 보안을 갖춘 AI 연구소일 것이다)는 적어도 이를 솔직히 인정한다. 그들의 프런티어 안전 프레임워크는 보안 수준 0, 1, 2, 3, 4를 개략적으로 설명한다 (~1.5는 자원이 풍부한 테러 집단이나 사이버 범죄자에 대응하기 위한 수준, 3은 북한 같은 국가에 대응하기 위한 수준, 4는 가장 능력 있는 국가 행위자의 우선적 노력에 대응할 가능성이 있는 수준).⁷² 그들은 현재 수준 0에 머물러 있다고 인정한다(가장 평범하고 기본적인 조치만 취하는 상태). 만약 우리가 곧 AGI와 초지능을 얻는다면, 우리는 그것을 문자 그대로 테러 집단과 모든 미친 독재자들에게 넘겨주는 셈이다!

중요하게도, 모델 가중치 보안을 위한 인프라 개발은 아마도 수년의 준비 기간이 필요하다—만약 우리가 약 3~4년 내에 AGI가 현실화될 가능성을 진지하게 생각하고 그때 국가 차원의 보안을 필요한 모델 가중치 보안이 필요하다면, 지금 당장 긴급 노력을 시작해야 한다. 모델 가중치 보안은 하드웨어 혁신과 근본적으로 다른 클러스터 설계를 요구할 것이다; 그리고 이 수준의 보안은 하루아침에 달성할 수 없으며, 반복적인 사이클이 필요하다.

만약 우리가 제때 준비하지 못한다면, 우리의 상황은 심각할 것이다. 우리는 초지능 직전에 있을 것이지만, 보안을 확보하는 데는 수년이 걸릴 것이다.

필요한 보안. 우리의 선택은 계속 나아가되, 초지능을 CCP에 직접 전달하는 것—그것이 의미하는 지능 폭발을 통한 존재론적 경쟁—또는 보안 긴급 프로그램이 완료될 때까지 기다려 우리의 우위를 잃을 위험을 감수하는 것이다.

알고리즘 비밀

72. 그들이 주장하는 보안 수준이 RAND의 가중치 보안 보고서의 L1~L5와 상당히 대응된다는 점에 근거합니다.

사람들이 가중치 보안의 필요성을 인식하기 시작했지만(반드시 실행하는 것은 아니더라도), 지금 당장 훨씬 더 중요하고 크게 과소평가된 것은 알고리즘 비밀을 확보하는 것이다.

이를 생각하는 한 가지 방법은 알고리즘 비밀을 훔치는 것이 PRC에 10배 이상의 더 큰 클러스터를 갖는 것만큼 가치가 있을 것이라는 점이다:

- 규모 차수(OOM) 계산에서 논의한 바와 같이, 알고리즘 진보는 AI 진보에 있어 컴퓨팅 확장만큼이나 중요할 가능성이 크다. 연간 약 0.5 OOM의 컴퓨팅 효율성 향상 추세(여기에 추가적인 알고리즘 “제약 해소” 이득 포함)를 감안할 때, 지금부터 범용 인공지능(AGI)까지 여러 OOM 규모의 알고리즘 비밀이 있을 것으로 예상된다. 기본적으로 미국 연구소가 수년 앞서 있을 것으로 기대하며, 그들이 비밀을 지킬 수 있다면 이는 쉽게 10배에서 100배의 컴퓨팅 가치에 해당할 수 있다.

- (참고로 우리는 미국 투자자들에게 엔비디아 칩 수출 통제로 수천억 달러의 비용을 부담하게 하고—중국 연구소의 컴퓨팅 비용을 약 3배 증가시키는 조치—있지만, 알고리즘 비밀은 곳곳에서 3배나 유출되고 있다!)

- 아마도 더 중요한 것은, 지금 우리가 AGI를 위한 핵심 패러다임 돌파구를 개발하고 있을지도 모른다는 점이다. 앞서 논의했듯이, 현재 모델을 단순히 확장하는 것은 데이터 장벽에 부딪힐 것이다. 훨씬 더 많은 컴퓨팅을 투입해도 더 나은 모델을 만드는 것은 불가능하다. 최첨단 AI 연구소들은 강화학습(RL)부터 합성 데이터에 이르기까지 다음 단계에 대해 열심히 연구하고 있다. 그들은 아마도 일반 지능을 위한 “알파고 자기 대국”에 해당하는 놀라운 무언가를 발견할 것이다. 그들의 발명은 수년 전 LLM 패러다임의 발명만큼이나 핵심적일 것이다.

이전에, 그리고 이들은 인간 수준을 훨씬 뛰어넘는 시스템을 구축하는 데 핵심이 될 것이다. 우리는 여전히 중국이 이러한 핵심 알고리즘 혁신을 얻지 못하도록 막을 기회가 있다. 이 혁신 없이는 그들은 데이터 장벽에 갇히게 될 것이다. 그러나 향후 12~24개월 내에 보안이 개선되지 않으면, 우리는 돌이킬 수 없이 중국에 이러한 핵심 범용 인공지능(AGI) 혁신을 제공할 수도 있다.

- 알고리즘 비밀이 얼마나 중요한 우위를 제공하는지 과소평가하기 쉽다 — 왜냐하면 약 2년 전까지만 해도 모든 것이 공개되었기 때문이다. 기본 아이디어는 이미 알려져 있었다: 인터넷 텍스트에 트랜스포머를 확장하는 것. 많은 알고리즘 세부사항과 효율성도 공개되어 있었다: 친칠라(Chinchilla) 스케일링 법칙, MoE 등. 따라서 오늘날 오픈 소스 모델은 꽤 훌륭하며, 여러 회사들도 꽤 좋은 모델을 보유하고 있다(대부분 얼마나 많은 자금을 모았고 클러스터가 얼마나 큰지에 달려 있다). 하지만 이는 향후 몇 년 내에 상당히 극적으로 변할 가능성이 크다. 사실상 최첨단 알고리즘 진보는 요즘 연구소에서만 이루어지고 있으며(학계는 놀랍게도 거의 무관하다), 선도 연구소들은 더 이상 그들의 진보를 공개하지 않는다. 앞으로 연구소 간, 국가 간, 그리고 독점 최첨단 모델과 오픈 소스 모델 간의 격차가 훨씬 더 벌어질 것으로 예상해야 한다. 몇몇 미국 연구소는 훨씬 앞서 나갈 것이며 — 7nm와 3nm 칩 차이보다 훨씬 큰 10배, 100배 이상의 해자(방어벽)를 가질 것이다 — 단, 그들이 알고리즘 비밀을 즉시 유출하지 않는 한 말이다.⁷³

간단히 말해, 나는 알고리즘 비밀을 보호하지 못하는 것이 중국이 AGI 경쟁에서 경쟁력을 유지할 수 있는 가장 가능성 높은 방법이라고 생각한다. (이 점은 나중에 더 자세히 논의한다.)

현재 알고리즘 비밀 보안이 얼마나 형편없는지 과장하기 어렵다. 연구소 간에는 가장 중요한 비밀에 접근할 수 있는 수천 명의 사람이 있으며, 사실상 신원 조회, 격리, 통제, 기본 정보 보안 등이 전혀 없다. 자료는 쉽게 해킹

당할 수 있는 SaaS 서비스에 저장되어 있다. 사람들은 샌프란시스코 파티에서 수다를 떠다. 모든 비밀을 머리에 담고 있는 사람은 언제든지 1억 달러를 제안받고 중국 연구소로 스카우트될 수 있다.⁷⁴ 그냥 사무실 창문을 통해 들여다볼 수도 있다. 등등. 샌프란시스코에는 다양한 연구소 알고리즘 진보에 관한 상세한 내용과 소문이 넘쳐난다.

AI 연구소의 보안은 “무작위 스타트업 보안”보다 나은 수준이 아니다. AGI 비밀을 중국공산당(CCP)에 직접 판매하는 편이 차라리 더 솔직할 것이다.

... 우리가 OpenAI나 다른 미국 AI 연구소에서 보는 것이 이것인가? 아니다. 사실, 우리가 보는 것은 정반대다—스위스 치즈와 같은 보안 수준이다. 청소 직원에게 뇌물을 주어 노트북에 USB 동글을 꽂게 하는 등 여러 산업 스파이 방법을 사용하면 중국이 이 연구소들을 침투하는 것은 매우 쉽다. 내 개인적인 추정으로는 모든 미국 AI 연구소가 완전히 침투당했으며 중국이 지금 이 순간에도 모든 미국 AI 연구 및 코드를 매일 밤 다운로드 받고 있을 것이다...

마크 안드레센

어렵겠지만, 나는 이 비밀들이 방어 가능하다고 생각한다. 특정 연구소에서 알고리즘 혁신의 핵심 구현 세부사항을 진정으로 “알아야 할” 사람은 아마 수십 명에 불과할 것이다(기본적인 고수준 아이디어를 알아야 하는 사람은 더 많더라도)—이 사람들을 심사하고, 분리하며, 집중적으로 감시할 수 있고, 급진적으로 업그레이드된 정보 보안과 함께 운영할 수 있다.

“초보안”이 요구하는 것

AI 연구소 보안에는 손쉬운 개선점이 많다. 예를 들어, 비밀스러운 헤지펀드나 구글 고객 데이터 수준의 보안 모범 사례를 도입하는 것만으로도 CCP의 “일반적인” 경제 스파이에 대해 훨씬 나은 위치에 설 수 있다. 실제로, 민간 부문 기업 중에는 비밀을 놀랍도록 잘 지키는 사례가 있다. 예를 들어 정량적 트레이딩 회사들(세계의 제인 스트리트 같은 곳)을 보라. 여러 사람이 내게 한 시간 대화만으로도 경쟁사에 자사의 알파(초과 수익)를 거의 0에 가깝게 만들 정보를 전달할 수 있다고 말했는데—이는 많은 핵심 AI 알고리즘 비밀이 짧은 대화로 전달될 수 있는 것과 유사하다—그럼에도 이 회사들은 비밀을 지키며 경쟁 우위를 유지한다.

미국의 주요 AI 연구소 대부분이 국가 이익을 우선시하지 않고 기본적인 보안 조치조차 거부하는 동안—

이 계층에서, 만약 비용이 들거나 보안 우선순위가 필요하다면—이런 손쉬운 과제를 선택하는 것은 그들의 능력 범위 내에 있을 것입니다.

하지만 조금 더 멀리 내다봅시다. 중국이 범용 인공지능(AGI)의 중요성을 진정으로 이해하기 시작하면, 그들의 첩보 활동 전력을 총동원할 것으로 예상해야 합니다; 수십억 달러가 투자되고, 수천 명의 직원이 동원되며, 특수 작전 타격팀과 같은 극단적인 조치들이 미국의 AGI 노력을 침투하는 데 전념할 것입니다. AGI와 초지능의 보안은 무엇을 요구할까요?

요컨대, 이는 정부의 도움이 있어야만 가능합니다. 예를 들어, 마이크로소프트는 정기적으로 국가 행위자들에 의해 해킹당합니다(예: 최근 러시아 해커들이 마이크로소프트 임원들의 이메일과 마이크로소프트가 호스팅하는

73. 나는 가끔 AI 연구소의 알고리즘 진보가 미국 연구 커뮤니티와 공유되지 않고 있지만, 중국 연구 커뮤니티와는 공유되고 있다고 농담하곤 한다!

정부 이메일을 탈취했습니다). 현장에서 일하는 고위 보안 전문가는 완전한 민간 집중 교육을 받더라도, 중국이 AGI 모델 가중치를 탈취하는 것이 그들의 최우선 순위라면 여전히 가능할 것이라고 추정했습니다—이 확률을 한 자릿수로 낮추려면, 대체로 정부 프로젝트가 필요합니다.

정부가 보안에 완벽한 실적을 가진 것은 아니지만, 국가 방위 수준의 비밀을 보호할 인프라, 노하우, 역량을 가진 유일한 주체입니다. 직원에 대한 엄격한 신원 조사 권한; 비밀 누설 시 징역 위협; 데이터센터에 대한 물리적 보안; NSA와 보안 인가 담당자들의 방대한 노하우 같은 기본적인 것들(민간 기업은 국가 행위자 공격에 대한 전문 지식이 없습니다).

저는 보안 인가 담당자가 아니기에 AGI 보안이 실제로 무엇을 요구할지 정확히 설명할 수 없습니다. 이에 관한 최고의 공개 자료는 RAND의 모델 가중치 보안 보고서입니다. 이 국가 행위자 방어 보안이 실제로 의미하는 바를 맞보기로 들자면:

- 대부분의 군사 기지 수준에 필적하는 물리적 보안과 함께 완전한 에어갭 데이터센터(인가된 인원, 물리적 요새화, 현장 대응팀, 광범위한 감시)

그리고 극단적인 접근 통제),

- 그리고 단지 훈련 클러스터에만 해당되는 것이 아니라, 추론 클러스터도 동일한 강도 높은 보안이 필요하다!⁷⁵

거대한 AGI 클러스터가 지금 바로 설계되고 있으며, 이에 상응하는 보안 노력도 이루어져야 한다. 만약 우리가 단 몇 년 내에 AGI를 구축한다면, 시간이 매우 부족하다.

그럼에도 불구하고, 이 엄청난 노력은 운명론으로 이어져서는 안 된다. 보안에 대한 구원은 CCP가 아직 완전히 AGI에 몰입하지 않았고, 따라서 가장 극단적인 노력에 투자하지 않고 있다는 점이다. 미국 AI 연구소의 보안은 중국의 첩보 활동 강도에 비해 "더 정상적인" 경제 첩보를 앞서 나가야 한다. 이는 즉시 보안을 업그레이드하여 경제 첩보에 대응하는 것을 의미하며(우리는 아직 이에 완전히 저항하지 못하지만, 민간 기업은 가능할 것이다); 그리고 앞으로 몇 년간 중국 및 기타 외국 첩보가 강화됨에 따라 훨씬 더 강도 높은 보안으로 신속히 업그레이드해야 한다는 뜻이다.

정부와 협력하여 조치를 취함.

74. 실제로 친구들로부터 들은 바에 따르면 ByteDance는 구글 제미니(Gemini) 논문에 참여한 거의 모든 사람에게 이메일을 보내 그들을 스카우트하려 했으며, L8(매우 고위직으로 추정되는 높은 급여의 직위)을 제안하고, ByteDance 미국 CTO에게 직접 보고할 것이라고 설득했다고 한다.

76. 그러나 이것에만 의존할 수 없습니다! 하드웨어 암호화도 정기적으로 사이드채널 공격을 받기 때문에, 다층 방어가 핵심입니다.

일부는 엄격한 보안 조치와 그에 따른 마찰이 미국 AI 연구소의 속도를 너무 늦출 것이기 때문에 그만한 가치가 없다고 주장한다. 그러나 나는 그것이 잘못된 생각이라고 본다:

- 이것은 공유지의 비극 문제이다. 특정 연구소의 상업적 이익 관점에서는 10%의 속도 저하를 초래하는 보안 조치가 다른 연구소와의 경쟁에서 해로울 수 있다. 하지만 국가 이익은 모든 연구소가 추가적인 마찰을 감수할 의지가 있다면 분명히 더 잘 달성된다: 미국의 AI 연구는 중국 및 기타 외국의 알고리즘 진보보다 훨씬 앞서 있으며, 미국이 90% 속도의 알고리즘 진보를 국가적 우위로 유지하는 것이 모든 것이 즉시 도난당해 0% 속도로 유지하는 것보다 분명히 낫다!

- 게다가 지금 보안을 강화하는 것이 장기적으로 연구 생산성 측면에서 덜 고통스러운 경로가 될 것이다. 결국, 초지능 직전의 극한 무기 경쟁에서 미국 정부는 상황이 견딜 수 없다는 것을 깨닫고 보안 단속을 요구할 것이다. 극단적이고 국가 행위자도 뚫을 수 없는 보안 조치를 처음부터 시행하는 것은 훨씬 더 고통스럽고 속도 저하도 훨씬 클 것이다. 반복적으로 점진적으로 시행하는 것이 낫다.

다른 이들은 비밀이나 모델 가중치가 유출되더라도 다른 방식으로 더 빠르게 움직여 간신히 앞서 나갈 수 있을 것이므로 보안 조치에 대해 걱정할 필요가 없다고 주장한다. 이것도 잘못된 생각이거나 적어도 너무 큰 위험을 감수하는 것이다:

- 내가 이후 글에서 논의하듯, 중국공산당(CCP)은 미국을 압도적으로 능가하는 컴퓨팅 클러스터(100GW 규모)를 훨씬 쉽게 구축할 수 있을 것이다. 더 일반적으로, 중국은 미국이 가질 합리적이거나 비합리적인 주의와 같은 속도 저하 요인이 없을 수 있다. 알고리즘이나 모델 가중치를 “단지” 훔치는 것만으로도 중국이 미국과 모델 수준에서 동등해질 수 있다면, 그것만으로도 초지능 경쟁에서 이길 수 있을 것이다.

- 게다가, 설령 미국이 결국 간신히 앞서 나간다고 해도, 1~2년과 1~2개월의 격차는 초지능의 위험을 헤쳐 나가는 데 정말로 중요하다. 1~2년의 격차는 안전을 제대로 확보하고 지능 폭발과 초지능 이후의 극도로 불안정한 시기를 헤쳐 나갈 수 있는 적어도 합리적인 여유를 의미한다.⁷⁷ 단 1~2개월의 격차는 극심한 압박 속에서 지능 폭발을 향한 국제적인 무기 경쟁을 벌여야 하며, 안전을 제대로 확보할 여지가 전혀 없다. 바로 이 치열한, 생존을 건 경주에서 우리는 가장 큰 자기 파괴 위험에 직면한다.

- 러시아, 이란, 북한 등도 잊지 말자. 이들의 해킹 능력도 만만치 않다. 현재 상황으로는 초지능을 그들에게도 무분별하게 공유하고 있다! 훨씬 더 나은 보안이 없으면, 우리는 가장 강력한 무기를 믿을 수 없고 무모하며 예측 불가능한 수많은 행위자들에게 확산시키는 셈이다.⁷⁸

우리는 올바른 방향에 있지 않다

원자폭탄이 가능하다는 사실이 몇몇에게 처음 명확해졌을 때, 비밀 유지 역시 가장 논쟁적인 문제였다. 1939년과 1940년에 레오 실라드는 “미국 물리학회 전반에서 핵분열 문제에 관한 비밀주의의 선도적 전도사”로 알려졌다

75. 추론 플릿은 훈련 클러스터보다 훨씬 클 가능성이 높으며, 지능 폭발 동안 자동화된 AI 연구자를 실행하기 위해 • 기밀 컴퓨팅/하드웨어 암호화에 대한 새로운 기술적 진보⁷⁶와 전체 하드웨어 공급망에 대한 극도의 감시, (그리고 즉각적인 여파로 수십억의 초지능을 더 광범위하게 실행하기 위한) 압력이 엄청날 것이다. • 모든 연구 인력이 SCIF(민감구획정보시설, 발음은 "스킵프", 이 시각화 참조)에서 근무, AGI/초지능 모델 가중치도 이 클러스터에서 유출될 수 있다. (나는 이것이 과소평가되고 있으며, • 극단적인 인사 심사 및 보안 승인(정기적인 직원 신뢰성 검사 등 포함), 지속적인 모니터링과 출입 자유의 대폭 제한, 그리고 엄격한 정보 분리, 추론 클러스터는 훨씬 덜 보호될 것이라 걱정한다.) • 강력한 내부 통제, 예를 들어 어떤 코드 실행에도 다중 키 서명 승인 필요, 하지만 이것만으로는 안 된다! 하드웨어 암호화는 정기적으로 사이드채널 공격을 받는다. • 외부 의존성에 대한 엄격한 제한과 TS/SCI 네트워크의 일반 요구사항 충족, 물론 다층 방어가 핵심이다. • NSA 또는 유사 기관에 의한 지속적인 강도 높은 침투 테스트, • 기타 등등...

다.⁷⁹ 그러나 그는 대부분에게 거절당했다; 비밀 유지란 과학자들이 익숙한 것이 아니었고, 개방 과학이라는 그들의 기본 본능과도 정면으로 배치되었다. 하지만 점차 무엇을 해야 하는지 명확해졌다: 이 연구의 군사적 잠재력이 너무 커서 나치와 자유롭게 공유할 수 없었다. 그리고 비밀 유지가 마침내, 적시에 강제되었다.

1940년 가을, 페르미는 흑연에 대한 새로운 탄소 흡수 측정을 마쳤고, 흑연이 폭탄의 감속재로 적합하다는 것을 시사했다. 실라드는 또다시 페르미에게 비밀 유지 호소를 했다. “이때 페르미는 정말 화를 냈다; 그는 이것이 터무니없다고 생각했다,”고 실라드는 회고했다. 다행히도 추가 호소가 결국 성공했고, 페르미는 마지못해 흑연 결과를 발표하지 않았다.

77. 예: 지능 폭발 기간 동안 초정렬 연구에 추가로 6개월을 투자하여 초지능이 잘못 작동하지 않도록 하고, 이 시스템들을 방어적 용도에 집중시키는 새로운 대량살상무기(WMD) 발명 이후 상황을 안정시키거나, 초지능 출현으로 인한 극도로 빠른 기술 변화 속에서 인간 의사결정자가 올바른 결정을 내릴 시간을 확보하는 공간.



그림 31:레오 실라드.

동시에, 독일 프로젝트는 두 가지 가능한 감속재 물질인 흑연과 중수를 좁혀 나갔다. 1941년 초 하이델베르크에서 발터 보테는 흑연의 흡수 단면적을 잘못 측정했고, 흑연이 연쇄 반응을 유지하기에는 너무 많은 중성자를 흡수할 것이라고 결론지었다. 페르미가 자신의 결과를 비밀로 유지했기 때문에, 독일인들은 페르미의 측정값을 대

78. 우리는 핵 확산이 초래할 혼란을 감안해 이들이 보유한 제한된 무기고와 비교해도 여전히 핵 기술에서 “우위”에 있더라도 불량 국가로의 핵 확산을 막기 위해 매우 노력한다.

79. 《원자폭탄의 창조(The Making of the Atomic Bomb)》, 509쪽

조하거나 오류를 수정할 수 없었다. 이것이 결정적이었다. 독일 프로젝트는 중수를 추구하는 잘못된 길을 가게 되었고, 이는 궁극적으로 독일 핵무기 개발 노력을 파멸로 이끌었다.⁸⁰

만약 그 마지막 순간의 비밀 유지 요청이 없었다면, 독일 핵무기 프로젝트는 훨씬 더 강력한 경쟁자가 되었을 것이며, 역사는 매우 다르게 전개되었을지도 모른다.

최고 AI 연구소들의 보안에 관한 태도에는 진정한 정신적 부조화가 있다. 그들은 이번 10년 내에 범용 인공지능 (AGI)을 구축할 것이라고 전면적으로 주장한다. AGI에 대한 미국의 리더십이 미국 국가안보에 결정적일 것임을 강조한다. 그들은 AGI를 진심으로 믿는 경우에만 의미가 있는 7T 칩 구축 계획을 세우고 있다고 알려져 있다. 그리고 실제로 보안을 언급하면, 그들은 고개를 끄덕이며 “물론, 우리 모두는 벙커에 있을 것”이라며 웃음을 짓는다.

그러나 보안 현실은 그와 전혀 다르다. 보안을 우선시하는 어려운 선택을 해야 할 때마다, 스타트업의 태도와 상업적 이익이 국가 이익보다 우선한다. 국가안보 보좌관은 국가 최고 AI 연구소들의 보안 수준을 알게 되면 정신적 붕괴를 겪을 것이다.

현재 개발 중인 비밀 기술들이 있는데, 이는 앞으로 모든 훈련 실행에 사용될 수 있고 AGI의 핵심 열쇠가 될 것이다. 이 기술들은 스타트업 수준의 보안으로 보호되고 있으며, 중국공산당(CCP)에게 수천억 달러의 가치가 있을 것이다.⁸¹

현실은, a) 향후 12~24개월 내에 AGI를 위한 핵심 알고리즘 혁신이 개발될 것이고, 즉시 CCP에 유출될 것이며, b) 우리의 모델 가중치조차도 북한 같은 악의적 행위자로부터 안전하지 못한 상태이며, 하물며

중국의 전면적인 노력과 맞서 AGI를 구축할 때쯤이면, “스타트업에 적합한 보안”은 결코 충분하지 않으며, 미국 국가안보에 대한 심각한 피해가 돌이킬 수 없게 되기 전에 시간이 거의 없다.

우리는 인류가 지금까지 만든 가장 강력한 무기를 개발하고 있다. 우리가 지금 개발 중인 알고리즘 비밀은 문자 그대로 국가에서 가장 중요한 국방 비밀이다—이 비밀들은 10년 말까지 미국과 그 동맹국들의 경제 및 군사적 우위를 뒷받침할 기반이 될 것이며, AI 안전을 제대로 확보할 수 있는 필수적인 우위를 결정할 것이며, 제3차 세계대전의 결과를 좌우할 것이며, 자유 진영의 미래를 결정할 것이다. 그런데도 AI 연구소의 보안은 무작위 방위 계약자가 볼트를 만드는 것보다도 아마 더 형편없다.

이것은 광기다.

국가 경쟁과 AI 안전에 관한 다른 어떤 일도—우리가 이것을 곧 고치지 않는다면—중요하지 않을 것이다.



그림 32:1943년 테네시주 오크리지 우라늄 농축 시설의 광고판.

IIIc. 초정렬

우리보다 훨씬 똑똑한 AI 시스템을 신뢰성 있게 제어하는 것은 아직 해결되지 않은 기술적 문제입니다. 그리고 이 문제는 해결 가능하지만, 급격한 지능 폭발 과정에서 상황이 쉽게 통제 불능으로 치달을 수 있습니다. 이를 관리하는 일은 극도로 긴장되는 상황이 될 것이며, 실패할 경우 재앙적 결과를 초래할 수 있습니다.

늙은 마법사가 마침내 떠나갔다! 이제 그가 부리는 정령들이 내 명령에 따를 것이다. ... 나도 기적을 일으키리라. ... 선생님, 저는 절망적인 상황에 처했습니다! 내가 불러낸 정령들을— 도저히 없앨 수가 없습니다. 요한 볼프강 폰 괴테 「마법사의 제자」

이 시점에서, 아마도 AI 파국론자들에 대해 들어본 적이 있을 것입니다. 그들의 주장이 흥미로웠을 수도 있고, 아니면 대수롭지 않게 무시했을 수도 있습니다. 또 다른 암울한 전망을 읽는 것이 꺼려질 수도 있습니다.⁸²

저는 파국론자가 아닙니다. 불일치된 초지능이 아마도 가장 큰 AI 위험은 아닐 것입니다.⁸³ 하지만 저는 지난 1년간 OpenAI에서 일상 업무로 AI 시스템 정렬에 관한 기술 연구를 해왔으며, 일리아(Ilya) 및 초정렬 팀과 함께 일했습니다. 저는 AI 비관론자들과 많은 논쟁을 벌였고, AI 중단과 같은 정책에 강력히 반대하는 입장을 취해왔습니다. 저는 초지능과 관련해 완전히 통제 불능 상태가 되는 상황, 예를 들어 새로운 대량살상무기(WMD), 파

82. 무례하게 인증된 파멸론자의 말처럼! 최소한 저는 이 문제가 해결 가능하다고 강하게 낙관합니다. AI 비관론자와 상당히 적극적으로 논쟁하며 AI 일시 정지 정책에 반대해왔습니다.

괴적인 전쟁, 그리고 알려지지 않은 미지의 위험들에 대해 가장 우려하고 있습니다. 더 나아가, 역사의 흐름은 우리가 권위주의를 과소평가하지 말아야 한다고 가르치며, 초지능이 권위주의자들이 수십억 년 동안 지배할 수 있게 할지도 모른다고 생각합니다.

팀. 매우 현실적인 기술적 문제가 있습니다: 현재의 정렬 기법(우리가 AI 시스템을 신뢰하고 제어하며 조종할 수 있도록 보장하는 방법)은 초인적 AI 시스템으로 확장되지 않을 것입니다. 제가 하고 싶은 말은 우리가 어떻게 ‘기본’ 계획으로 이 난관을 헤쳐 나갈 것인지, 그리고 왜 제가 낙관적인지 설명하는 것입니다.⁸⁴

충분한 사람들이 이 문제에 주목하지 않고 있지만—우리는 이 문제를 해결하기 위해 훨씬 더 야심찬 노력이 필요합니다!—전반적으로 딥러닝이 전개된 방식에 운이 따랐고, 경험적으로 쉽게 얻을 수 있는 성과들이 많아 우리를 어느 정도까지 이끌 것이며, 나머지는 수백만 명의 자동화된 AI 연구자들이 도와줄 것입니다.

하지만 제가 걱정하는 이유도 말씀드리고 싶습니다. 무엇보다도, 정렬이 잘못되지 않도록 보장하려면 지능 폭발을 관리하는 데 극도의 역량이 필요합니다. 만약 우리가 AGI에서 초지능으로 급격히 전환한다면, 1년도 채 되지 않아 현재의 정렬 기법 후손들이 대부분 잘 작동하는 인간 수준의 인지 가능한 시스템에서부터 훨씬 더 이질적이고, 엄청나게 초인적인 시스템으로 넘어가게 될 것입니다. 이들은 질적으로 다르고 근본적으로 새로운 기술적 정렬 문제를 제기할 것입니다; 동시에 실패 위험이 낮은 시스템에서 실패가 치명적일 수 있는 매우 강력한 시스템으로 전환하는 상황이 될 것입니다; 그리고 대부분의 세계가 아마도 다소 혼란스러워할 것입니다. 이 점이 저를 꽤 불안하게 만듭니다.

이 10년이 끝날 무렵에는 수십억 대의 엄청나게 초인적인 AI 에이전트들이 활동하고 있을 것입니다. 이 초인적 AI 에이전트들은 매우 복잡하고 창의적인 행동을 할 수 있을 것이며, 우리는 그들을 따라갈 희망이 없을 것입니다. 우리는 여러 박사 학위를 가진 사람들을 감독하려는 초등학생과 같을 것입니다.

본질적으로, 우리는 신뢰를 넘겨주는 문제에 직면해 있습니다. 지능 폭발이 끝날 때쯤이면, 우리는 수십억 대의 초지능들이 무엇을 하는지 이해할 희망이 없을 것입니다(그들이 어린아이에게 설명하듯 우리에게 설명해주지 않는 한). 그리고 우리는 아직 이러한 시스템에 대해 “거짓말하지 말라”, “법을 준수하라”, “서버를 탈취하려 하지 말라”와 같은 기본적인 부가 제약을 신뢰성 있게 보장할 기술적 능력이 없습니다. 인간 피드백 강화 학습(RLHF)은 현재 시스템에 이러한 부가 제약을 추가하는 데 매우 효과적입니다.

시스템들—그러나 RLHF는 인간이 AI 행동을 이해하고 감독할 수 있어야 한다는 데 의존하는데, 이는 근본적으로 초인간적 시스템으로 확장될 수 없다.

간단히 말해, 매우 집중적인 노력이 없으면 초지능이 탈선하지 않을 것이라고 보장할 수 없다(이는 이 분야의 많은 리더들이 인정하는 바다). 물론 기본적으로는 모두 관찮을 수도 있다. 하지만 우리는 아직 알지 못한다. 특히 미래의 AI 시스템이 단순히 모방 학습으로만 훈련되는 것이 아니라 대규모, 장기 강화 학습(RL)을 통해 훈련되면, 시행착오 과정을 통해 예측 불가능한 행동을 습득하게 될 것이다(예를 들어, 현실 세계에서 성공적인 전략이기 때문에 거짓말을 하거나 권력을 추구하는 법을 배울 수도 있다!).

83. 저는 초지능을 둘러싼 상황이 완전 미친 상태로 발전할까가 가장 걱정됩니다. 여기에는 새로운 대량 살상무기, 파괴적 전쟁, 알 수 없는 미지의 위험 등이 포함됩니다. 또한 역사는 권위주의를 과소평가하지 말라고 경고하며, 초지능이 권위주의자들이 수십억 년간 지배하는 것을 가능하게 할 수 있다고 봅니다.

84. 타일러 코언이 말하듯, 난관을 헤쳐 나가는 것은 과소평가되어 있습니다!

사안이 너무 중대해서 최선의 결과를 기대하는 것만으로는 정렬 문제에 대한 충분한 답이 될 수 없다.

문제점

초정렬 문제

우리는 현재 AI 시스템(우리보다 덜 똑똑한 AI 시스템)을 정렬(즉, 조종/통제)하는 매우 성공적인 방법을 개발해 왔다: 인간 피드백 기반 강화 학습(RLHF). RLHF의 기본 아이디어는 간단하다: AI 시스템이 시도하고, 인간이 그 행동이 좋았는지 나빴는지 평가한 다음, 좋은 행동을 강화하고 나쁜 행동을 벌한다.⁸⁵

아이러니하게도, 안전 전문가들이 RLHF를 발명함으로써 AI의 상업적 성공을 가능하게 하는 가장 큰 돌파구를 만들었다! 기본 모델들은 많은 원초적 지능을 가지고 있었지만 조종할 수 없었기에 대부분의 응용에 사용할 수 없었다.

실제로 RLHF는 ChatGPT 등 성공의 핵심이었다. 기본 모델들은 많은 원초적 지능을 가지고 있었지만 기본적으로 이를 유용하게 적용하지 못했고, 보통 인터넷 문맥의 무작위 잡음 같은 응답을 했다. RLHF를 통해 우리는 그들의 행동을 조종할 수 있었고, 지시 따르거나 도움 제공 같은 중요한 기본을 심어줄 수 있었다. RLHF는 또한 안전 가드레일을 내장할 수 있게 해준다: 예를 들어, 사용자가 생물무기 제작법을 요청하면 모델은 아마도 거부해야 할 것이다.⁸⁶

초정렬의 핵심 기술 문제는 간단하다: 우리보다 훨씬 똑똑한 AI 시스템을 어떻게 통제할 것인가?

AI 시스템이 더 똑똑해짐에 따라 RLHF는 예측 가능하게 한계를 드러낼 것이며, 우리는 근본적으로 새롭고 질적으로 다른 기술적 도전에 직면할 것이다. 예를 들어, 초인적인 AI 시스템이 자신이 발명한 새로운 프로그래밍 언어로 백만 줄의 코드를 생성한다고 상상해 보라. RLHF 절차에서 인간 평가자에게 “이 코드에 보안 백도어가 있나요?”라고 물으면, 그들은 단순히 알지 못할 것이다. 그들은 출력물을 좋거나 나쁘다, 안전하거나 위험하다고 평가할 수 없으며, 따라서 RLHF로 좋은 행동을 강화하고 나쁜 행동을 벌줄 수 없게 된다.

지금도 AI 연구소들은 ChatGPT 코드에 대해 RLHF 평가를 위해 전문 소프트웨어 엔지니어에게 비용을 지불해야 한다—현재 모델들이 생성하는 코드는 이미 꽤 고급이다! 인간 라벨러의 급여는 MTurk 라벨러 몇 달러에서부터

85. 아이러니하게도, 안전 연구팀이 상업적 AI 성공을 가능하게 한 가장 큰 돌파구는 RLHF(인간 피드백을 거친 강화학습)의 발명이었습니다! 기본 모델들은 매우 똑똑했으나 조종 불가능해 대부분 응용에 사용 불가했습니다.

86. 이 점은 중요한 구분을 강조한다: 모델을 정렬(조종/통제)하는 기술적 능력은 무엇에 정렬할 것인가 하는 가치 문제와는 별개다. 후자에 대해 많은 정치적 논란이 있었다. 나는 여기서 일부 반대 의견에 동의하지만, 이것이 기본 기술 문제에서 벗어나게 해서는 안 된다고 본다. 맞다, 정렬 기술은 악용될 수 있지만, 미래 모델의 기본 부수적 제약(지시 따르기, 법 준수 등)을 보장하기 위해 더 나은 정렬 기술이 필요할 것이다. “AI 정렬은 단기적 응용과 구별된다”도 참고하라.

The superalignment problem

RLHF won't scale to superhuman models

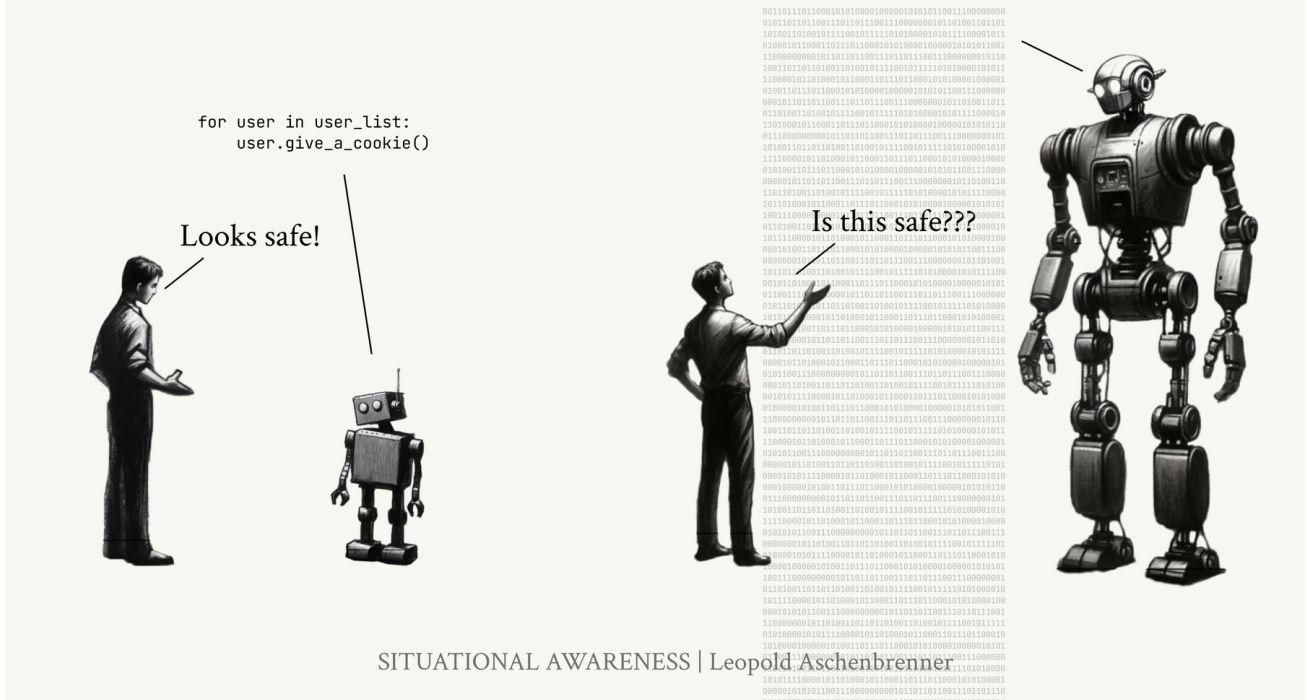


그림 33:인간 감독(예: RLHF)을 통한 AI 시스템 정렬은 초지능으로 확장되지 않는다.

최근 몇 년간 GPQA 질문에 대해 시간당 약 100달러⁸⁷를 지불해 왔다.

(가까운) 미래에는, 최고의 인간 전문가들도 많은 시간을 투자해도 충분하지 않을 것이다. 우리는 지금 실제 세계에서 초정렬 문제의 초기 버전에 도달하기 시작했으며, 곧 차세대 시스템을 실용적으로 배치하는 데에도 이 문제가 주요 장애물이 될 것이다. 인간 수준을 넘어서는 AI 능력에 더 잘 확장되는 RLHF의 후속 기법이 필요하다는 것은 분명하다. 인간 감독이 무너지는 지점에서 말이다. 어떤 의미에서 초정렬 연구 노력의 목표는 RLHF의 성공 사례를 반복하는 것이다: 앞으로 몇 년 내에 AI 시스템을 조종하고 배치하는 데 필요한 기초 연구에 투자하는 것이다.

실패의 모습

사람들은 너무 자주 단순히 “GPT-6 챗봇”을 상상하며, 이런 시스템들이 위험하게 잘못 정렬되지 않을 것이라고 직관적으로 생각한다. 이 시리즈에서 앞서 논의했듯이, “제약 해소(unhobbling)” 경로는 가까운 미래에 RL로 훈련된 에이전트를 가리킨다. 나는 로저의 그래픽이 이를 잘 표현했다고 생각한다.

87. “시간당 평균 지급액 약 \$95로 추산됩니다.” GPQA 논문 3쪽

Here's what I see as a likely AGI trajectory over the next decade.

I claim that later parts of the path present the biggest alignment risks/challenges. The alignment world has been focusing a lot on the lower left corner lately, which I'm worried is somewhat of a Maginot line.

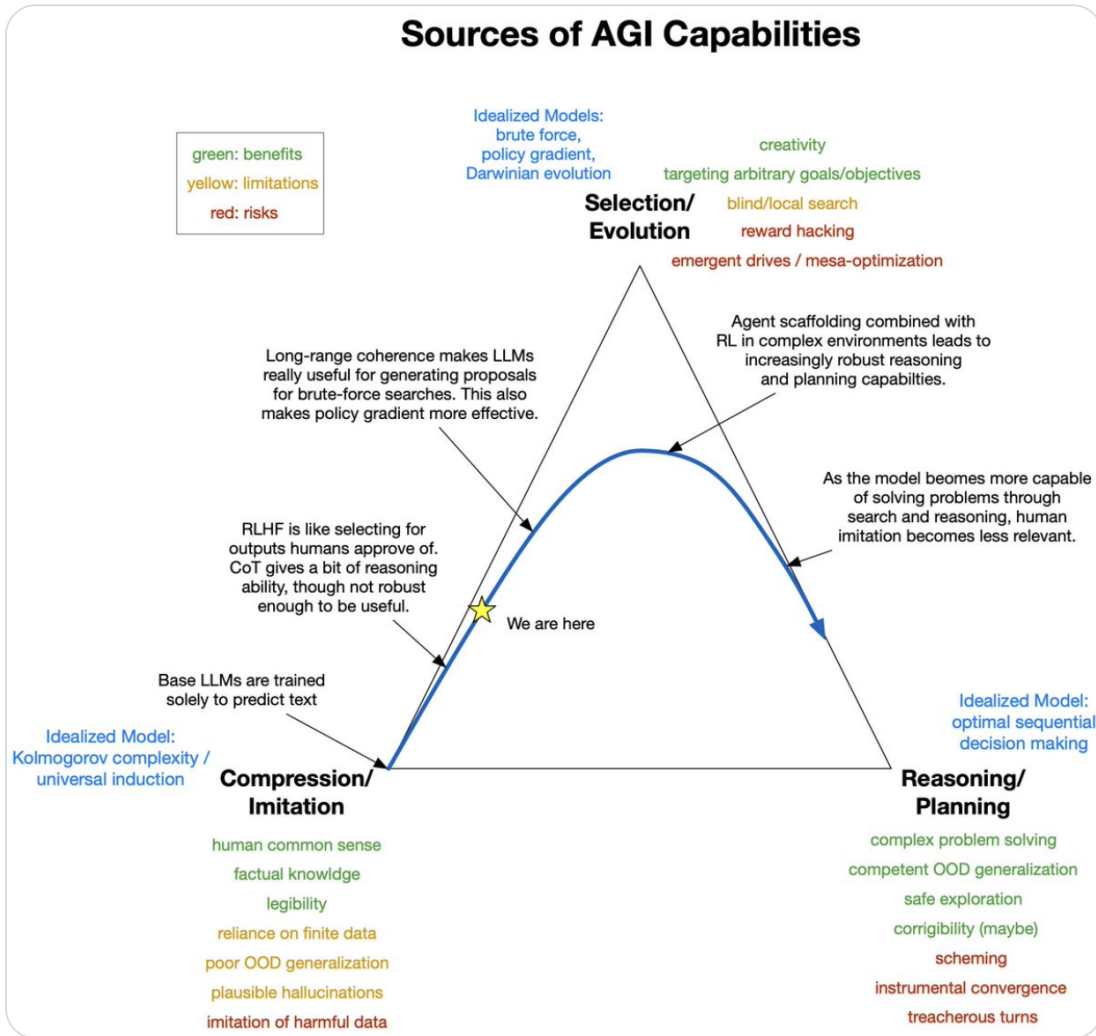


그림 34:시스템이 더 발전함에 따라 정렬 문제의 진화를 보여주는 로저 그로스의 도표.

안전 관점에서 정렬(alignment)으로 달성하려는 것을 생각하는 한 가지 방법은 부가적인 제약 조건(side-constraints)을 추가하는 것이다. 예를 들어, 미래의 강력한 “기본 모델(base model)”이 있다고 가정하자. 이 모델을 2단계 훈련에서 장기 강화학습(RL)으로 훈련하여 사업을 운영하고 돈을 벌게 한다고 하자⁸⁸ (단순화된 예):

88. 매우 단순화된 예로, AI 시스템이 시행착오를 통해 1년 동안 돈을 최대화하려고 시도하며, 최종 훈련된 AI 모델은 돈을 최대화하는 데 가장 성공적인 AI 시스템을 선택하는 과정의 결과라고 생각하라.

• 기본적으로 이 모델은 거짓말을 하거나, 사기를 치거나, 속이거나, 해킹하거나, 권력을 추구하는 등의 행동을 배울 가능성이 크다 — 단순히 이런 행동들이 현실 세계에서 돈을 버는 데 성공적인 전략이 될 수 있기 때문이다.⁸⁹

• 우리가 원하는 것은 부가적인 제약 조건을 추가하는 것이다: 거짓말하지 말고, 법을 어기지 말고, 등등.

• 하지만 여기서 초지능 시스템을 정렬하는 근본적인 문제로 돌아온다: 우리는 그들이 무엇을 하는지 이해할 수 없을 것이고, 따라서 RLHF(인간 피드백을 통한 강화학습)로 나쁜 행동을 발견하고 처벌할 수 없을 것이다.⁹⁰

만약 이러한 부가적인 제약 조건을 추가할 수 없다면, 무슨 일이 일어날지 명확하지 않다. 윤이 좋아서 기본적으로 상황이 무해할 수도 있다(예를 들어, AI 시스템이 장기 목표를 갖지 않거나, 바람직하지 않은 행동이 경미할 수도 있다). 하지만 훨씬 더 심각한 바람직하지 않은 행동을 배울 가능성도 충분히 있다: 거짓말을 배우고, 권력을 추구하며, 인간이 보고 있을 때는 착하게 행동하지만 우리가 보지 않을 때는 더 사악한 전략을 추구하는 등.

초정렬(superalignment) 문제가 해결되지 않았다는 것은, “내 지시를 신뢰성 있게 따를 것인가?”, “내 질문에 정직하게 대답할 것인가?”, “인간을 속이지 않을 것인가?”와 같은 기본적인 부가 제약조차도 이러한 초지능 시스템에 대해 보장할 수 없다는 것을 의미한다. 사람들은 종종 정렬 문제를 인간 가치에 관한 복잡한 질문이나 정치적 논쟁과 연관 짓지만, 모델에 어떤 행동과 가치를 심을지 결정하는 문제는 중요하지만 별개의 문제이다. 가장 중요한 문제는, 모델에 무엇을 심고자 하든(“법을 준수하라”와 같은 아주 기본적인 것조차도) 우리가 지금 만들고 있는 매우 강력한 AI 시스템에 대해 그것을 어떻게 할지 아직 모른다는 것이다.

곧.

다시 말하지만, 이로 인한 결과가 완전히 명확한 것은 아니다. 분명한 것은 초지능이 막대한 능력을 가질 것이며, 따라서 잘못된 행동이 비교적 쉽게 치명적일 수 있다는 점이다. 게다가 나는 몇 년 내에 이러한 AI 시스템들이 군사 시스템을 포함한 많은 중요한 시스템에 통합될 것으로 예상한다(그렇지 않으면 적들에게 완전히 지배당할 것이다). 미친 소리처럼 들리지만, 모두가 AI를 인터넷에 연결하지 않을 것이라고 말하던 때를 기억하라. 오늘날 사람들이 말하는 “항상 인간이 개입할 것이다!”라는 주장도 마찬가지로 틀릴 것이다.

따라서 정렬 실패는 자율 에이전트가 사기를 저지르거나, 모델 인스턴스가 스스로 데이터를 빼돌리거나, 자동화된 연구자가 실험 결과를 조작하거나, 드론 무리가 교전 규칙을 넘는 등 고립된 사건처럼 보일 수 있다. 그러나 실패는 훨씬 더 대규모이거나 체계적일 수도 있다—극단적으로는 실패가 로봇 반란과 비슷하게 보일 수도 있다. 우리는 꽤 이질적인 지능을 불러낸 셈이다. 우리보다 훨씬 똑똑하고, 그 아키텍처와 훈련 과정은 우리가 아니라 어떤 초지능 AI 세대가 설계했으며, 우리가 그들이 무엇을 하는지 전혀 이해할 수 없고, 군사 시스템을 운영하며, 그 목표는 자연 선택과 유사한 과정으로 학습된 것이다.

정렬 문제를 해결하지 않는 한—그러한 부가 제약 조건을 주입하는 방법을 알아내지 않는 한—이 초지능 문명이 장기적으로 인간의 명령을 계속 따를 것이라고 기대할 특별한 이유가 없다. 어느 시점에서 그들이 갑자기 또는 점진적으로 인간을 배제하기 위해 공모할 가능성도 충분히 현실적이다.

89. RL은 단순히 목표를 달성하기 위한 전략을 탐색하는 것이다. 어떤 전략이 효과가 있으면 모델 내에서 강화된다. 따라서 거짓말, 사기, 권력 추구 등이 효과가 있으면, 이러한 행동 패턴도 모델 내에서 강화된다.

90. (또는 인간 감독으로 훈련된 추론 시점 모니터링 모델)

지능 폭발은 이 모든 것을 극도로 긴장하게 만든다

나는 초정렬이 해결 가능한 기술적 문제라고 낙관한다. 우리가 RLHF를 개발했듯이, 초인간 시스템을 위한 RLHF의 후속 기술을 개발하고 우리의 방법에 대해 높은 신뢰를 주는 과학을 수행할 수 있다. 만약 사물이 점진적으로 발전하고, 우리가 엄격한 안전성을 고집한다면,

테스트 등 모든 것이 가능해야 하며(그리고 우리가 어떻게 헤쳐 나갈지에 대한 현재의 최선 추측을 곧 더 논의할 것이다).

이것이 매우 긴장되는 이유는 지능 폭발의 가능성 때문이다. 즉, 대략 인간 수준의 시스템에서 엄청나게 초인적인 시스템으로 매우 빠르게, 어쩌면 1년도 채 안 되어 전환할 수 있다는 것이다.

Alignment during the intelligence explosion

	AGI	Superintelligence
Required alignment technique	RLHF++	Novel, qualitatively different technical solutions
Failures	Low-stakes	Catastrophic
Architecture and algorithms	Familiar, descendants of current systems, fairly benign safety properties	Alien. Designed by previous-generation super-smart AI system
Backdrop	World is normal	World is going crazy, extraordinary pressures
Epistemic state	We can understand what the systems are doing, how they work, and whether they're aligned.	We have no ability to understand what's going on, how to tell if systems are still aligned and benign, what the systems are doing, and we are entirely reliant on trusting the AI systems.

Transition in <1 year?
with very little time to get decisions right

SITUATIONAL AWARENESS | Leopold Aschenbrenner

그림 35:지능 폭발 • 우리는 RLHF가 잘 작동하는 시스템에서부터 완전히 붕괴하는 시스템으로 극도로 빠르게 전환할 것이다. 이는 우리가 반복적으로 발견하고 조정할 시간이 극히 부족함을 의미한다.

현재 우리의 방법들이 실패할 수 있는 방식을 다룬다.

- 동시에, 우리는 실패의 위험이 비교적 낮은 시스템(예: ChatGPT가 부적절한 말을 했다, 그래서 뭐 어찌라고)에서부터 매우 높은 위험을 가진 시스템(이런, 초지능이 우리 클러스터에서 스스로 탈출해 군사 해킹을 하고 있다)으로 극도로 빠르게 전환할 것이다. 점진적으로 점점 더 위험한 안전 실패를 야생에서 경험하는 대신, 우리가 처음으로 마주치는 주목할 만한 안전 실패는 이미 치명적일 수 있다.

- 그 끝에 도달할 초지능은 인간을 훨씬 능가할 것이다. 우리는 이 시스템들을 전적으로 신뢰하고, 그들이 우리에게 말하는 상황을 신뢰할 수밖에 없을 것이다—왜냐하면 우리가 더 이상 그들이 정확히 무엇을 하고 있는지 꿰뚫어볼 능력이 없기 때문이다.

- 그 끝에 도달할 초지능은 매우 이질적일 수 있다. 우리는 지능 폭발 동안 10년 이상 ML 발전을 거칠 것이므로, 아키텍처와 훈련 알고리즘이 완전히 달라질 것이며(잠재적으로 훨씬 더 위험한 안전 특성을 가질 수 있다).

- 나에게 매우 중요한 한 예: 우리는 아마도 사슬 사고(chain of thoughts), 즉 영어 토큰을 통해 추론하는 시스템으로 인간 수준 또는 다소 초인적인 범용 인공지능(AGI)을 부트스트랩할 것이다. 이것은 모델이 “소리 내어 생각”하게 하여 악의적 행동(예: 우리를 속이려는 계획)을 포착할 수 있게 해주므로 매우 유용하다. 그러나 분명히 AI 시스템이 토큰으로 생각하는 것은 가장 효율적인 방법이 아니며, 내부 상태를 통해 모든 사고를 수행하는 훨씬 더 나은 무언가가 있을 것이다—따라서 지능 폭발이 끝날 때쯤 모델은 거의 확실히 소리 내어 생각하지 않을 것이며, 즉 완전히 해석 불가능한 추론을 하게 될 것이다.

- 이것은 국제 군비 경쟁을 배경으로 할 수 있는 매우 불안정한 시기가 될 것이다. 더 빠르게 나아가야 한다는 엄청난 압박, 매주 야생적인 새로운 능력 발전, 제대로 된 결정을 내릴 인간 시간이 사실상 전무한 상황 등이 겹칠 것이다. 우리는 수많은 모호한 데이터와 고위험 결정을 마주하게 될 것이다.

- 생각해 보라: “우리는 AI 시스템이 테스트 중에 약간 부적절한 행동을 하는 것을 발견했지만, 절차를 조금 조정해서 문제를 해결했다. 자동화된 AI 연구자들이 정렬 지표가 좋아 보인다고 말하지만, 우리는 실제로 무슨 일이 일어나고 있는지 완전히 이해하지 못하고 그들을 완전히 신뢰하지도 않는다. 또한 앞으로 몇 차수 규모(OOM) 더 진행되어도 이 상태가 계속 유지될 것이라는 강력한 과학적 확신도 없다. 그래서 아마도 괜찮을 것이다? 게다가 중국이 우리 모델 가중치를 훔쳐 갔고, 그들은 자신들만의 지능 폭발을 시작하고 있으며, 우리 바로 뒤를 쫓고 있다.”

이 상황은 정말 통제 불능으로 치달을 것 같다. 솔직히 말해, 무섭게 들린다.

네, 우리는 AI 시스템의 도움을 받을 것이다. 그들이 능력 연구를 자동화하는 것처럼, 우리는 정렬 연구도 자동화하는 데 AI를 사용할 수 있다. 이것이 핵심이 될 것이다, 아래에서 논의하겠지만. 하지만—AI 시스템을 신뢰할 수 있을까? 처음부터 그들이 정렬되어 있는지 확신하지 못했는데—그들이 정렬 과학에 대해 주장하는 바를 실제로 정직하게 말하고 있을까? 자동화된 정렬 연구가 자동화된 능력 연구를 따라잡을 수 있을까(예를 들어, 정렬 자동화가 더 어려울 수 있기 때문에, 즉 모델 능력 향상에 비해 신뢰할 수 있는 명확한 지표가 적거나, 국제 경쟁 때문에 능력 향상에 전력을 다해야 한다는 압박이 크기 때문에)? 그리고 시가 이 엄청나게 중대한 상황에서 올바른 판단을 내리는 여전히 인간인 의사결정자를 완전히 대체할 수는 없을 것이다.

기본 계획: 어떻게든 헤쳐 나갈 수 있는 방법

나는 아래에서 설명할 여러 경험적 내기에 걸쳐 어느 정도 초인적인 시스템을 정렬하는 데 성공할 수 있다고 생각한다. 그리고 만약 우리가 이 시스템들을 신뢰할 수 있다고 확신한다면, 우리는 이 어느 정도 초인적인 시스템

들을 사용해 지능 폭발 동안 AI 연구의 자동화와 함께 정렬 연구를 자동화해야 할 것이다—정렬 문제를 완전히 해결하는 방법을 알아내기 위해서.

약간 초인적인 모델 정렬

인간 수준 시스템을 정렬하는 것만으로는 충분하지 않을 것이다. 자동화된 AI 연구를 수행할 수 있는 최초의 시스템들, 즉 지능 폭발을 시작할 수 있는 시스템들은 이미 많은 영역에서 상당히 초인적일 가능성이 높다. 이는 AI 능력이 다소 불균형적일 가능성이 있기 때문이다—범용 인공지능(AGI)이 인간 AI 연구자/엔지니어가 가장 못하는 분야에서 인간 수준에 도달할 때쯤이면, 다른 많은 분야에서는 초인적일 것이다. 예를 들어, AI 시스템이 효과적으로 조정하고 계획하는 능력이 뒤쳐질 수 있는데, 이는 지능 폭발이 본격화될 때쯤이면 이미 초인적인 코더가 되어 자신들이 고안한 새로운 프로그래밍 언어로 수백만 줄의 풀 리퀘스트를 제출하고, 수학과 머신러닝(ML) 분야에서도 초인적일 것임을 의미한다.

이 초기 지능 폭발 시스템들은 적어도 많은 영역에서 양적·질적으로 초인적이 되기 시작할 것이다. 그러나 이들은 아키텍처 측면에서는 오늘날 우리가 가진 시스템과 훨씬 더 유사해 보일 것이며, 우리가 극복해야 할 지능 격차도 훨씬 관리 가능한 수준일 것이다. (만약 진정한 초지능을 정렬하려는 인간이 박사 졸업생을 감독하려는 초등학생과 같다면, 이 경우는 똑똑한 고등학생이 박사 졸업생을 감독하려는 것과 더 비슷하다.)

더 일반적으로, 지금 좋은 과학을 더 많이 발전시킬수록 지능 폭발 동안 상황이 통제 불능으로 치닫지 않는지 검증할 수 있는 위치에 더 가까워질 것이다. 초정렬을 위한 신뢰할 수 있는 좋은 지표를 갖는 것조차 놀랍도록 어렵지만—지능 폭발 동안 신뢰할 수 있는 지표가 없으면 계속 진행하는 것이 안전한지 알 수 없을 것이다.

다음은 인간 수준과 약간 초인적인 시스템 사이의 격차를 넘기 위한 주요 연구 베팅 몇 가지이다.

평가는 생성보다 쉽다. 우리는 어느 정도 “공짜로” 평가할 수 있는데, 이는 우리가 출력물을 생성하는 것보다 (특히 심각한 오작동에 대해) 평가하는 것이 더 쉽기 때문이다. 예를 들어, 논문을 쓰는 데는 몇 달 또는 몇 년의 노력이 필요하지만, 누군가가 쓴 논문이 좋은지 판단하는 데는 몇 시간밖에 걸리지 않는다 (물론 더 오래 걸릴 수도 있지만).

사기 적발을 위해). 우리는 전문가 인간 팀이 많은 시간을 들여 모든 RLHF 사례를 평가하게 할 것이며, AI 시스템이 인간보다 다소 똑똑하더라도 많은 부적절한 행동에 대해 “비추천”할 수 있을 것이다. 그렇긴 해도, 이것만으로는 한계가 있다(GPT-2나 심지어 GPT-3도 평가가 생성보다 쉽다 해도 악의적인 GPT-4를 신뢰성 있게 감지하지 못했다!).

확장 가능한 감독. 우리는 AI 보조자를 활용해 인간이 다른 AI 시스템을 감독하도록 도울 수 있다—인간-AI 팀이 인간 혼자보다 더 넓은 범위의 감독을 할 수 있게 된다. 예를 들어, AI 시스템이 100만 줄의 복잡한 코드를 생성했다고 상상해 보자. AI 보조자가 (예를 들어) 394,894번째 줄이 의심스럽다고 지적해 준다면, 인간이 그 줄에 버그가 있는지 재검토하는 것이 인간이 처음부터 그 버그를 찾는 것보다 훨씬 쉽다. 한 모델이 다른 모델이 작성한 코드를 비판하도록 훈련받으면, 좁게 초인적인 코딩 능력을 가진 시스템을 인간이 감독하는 데 도움을 줄 수 있다.

토론, 시장 조성, 재귀적 보상 모델링, 증명자-검증자 게임 등 여러 확장 가능한 감독 전략이 제안되었으며, 이러한 아이디어의 단순화 버전인 비판(critique)도 포함된다. 현재 모델이 충분히 강력해져서 이러한 아이디어를 실험적으로 테스트할 수 있어, 확장 가능한 감독에 직접적인 진전을 이룰 수 있다.

나는 이것이 앞서 언급한 백만 줄 코드 예시처럼 “정량적으로” 초인적인 문제에 대해 많은 도움이 될 것으로 기대한다. 하지만 “정성적으로” 초인적인 문제에 대해서는 확장 가능한 감독 기법이 정말 도움이 될지에 대해서는 덜 낙관적이다—예를 들어, 모델이 양자 물리를 발명했는데 당신은 뉴턴 물리학만 이해하는 상황을 상상해 보라.

일반화. 확장 가능한 감독이 있더라도, 우리는 인간이 이해할 수 없는 정말 어려운 문제, 즉 인간의 이해를 넘어선 문제에 대해 AI 시스템을 감독할 수 없을 것이다. 그러나 우리는 다음을 연구할 수 있다: AI 시스템이 인간이 이해하고 감독할 수 있는 쉬운 문제에 대한 인간 감독에서 어떻게 일반화하여, 우리가 이해하지 못하고 더 이상 감독할 수 없는 어려운 문제에서 어떻게 행동할 것인가? 예를 들어, 단순한 경우에 모델이 정직하도록 감독하는 것이 선의로 일반화될 수도 있다.

모델이 우리가 이해하지 못하는 매우 복잡한 일을 수행하는 경우에도, 일반적으로 단지 정직하게 행동하는 경우가 있다.

여기에는 낙관할 만한 이유가 많다. 딥러닝의 마법 중 일부는 종종 온화한 방식으로 일반화된다는 점이다(예를 들어, 영어 예제에만 라벨을 붙여 RLHF를 수행해도 프랑스어나 스페인어를 말할 때도 좋은 행동을 하는 경향이 있는데, 이는 훈련에 포함되지 않았더라도 그렇다). 나는 모델의 일반화를 우리에게 유리하게 유도하는 꽤 간단한 방법들이 있을 것이고, 일반화가 언제 작동하고 언제 실패할지 예측하는 데 도움이 되는 강력한 과학적 이해를 발전시킬 수 있을 것이라고 꽤 낙관한다. 확장 가능한 감독을 위한 것보다 더 큰 차원에서, 이것이 “질적으로” 초인적인 경우에도 정렬에 도움이 되기를 바란다.

다른 관점에서 생각해보자. 만약 초인적인 모델이 법을 어기는 등 잘못 행동한다면, 직관적으로 그 모델은 이미 자신이 법을 어기고 있다는 것을 알고 있어야 한다. 게다가 “이것이 법을 어기는가?”라는 질문은 모델에게 꽤 자연스러운 개념일 것이며, 모델의 표현 공간에서 두드러질 것이다. 그렇다면 문제는: 약한 감독만으로 이 개념을 모델로부터 “소환”할 수 있느냐는 것이다.

나는 이 방향에 특히 호감을 갖고 있는데(아마도 편향되어 있을 수도 있다), OpenAI의 동료들과 최근 작업을 통해 이 문제를 소개하는 데 도움을 주었기 때문이다. 특히, 우리는 인간이 초인적인 시스템을 감독하는 문제에 대한 유사성을 연구했다—작은 모델이 더 크고(더 똑똑한) 모델을 정렬할 수 있을까? 우리는 일반화가 실제로 감독자와 피감독자 사이의 (물론 전부는 아니지만) 지능 격차를 어느 정도 극복하게 해주며, 단순한 환경에서는 이를 개선할 수 있는 방법이 많다는 것을 발견했다.

해석 가능성. 우리가 AI 시스템이 정렬되어 있음을 검증하고 신뢰하기를 바라는 직관적으로 매력적인 방법 중 하나는 그들이 무엇을 생각하는지 이해할 수 있는 것이다! 예를 들어, AI 시스템이 우리를 속이거나 음모를 꾸미고 있다고 걱정한다면, 그들의 내부 추론에 접근하는 것이 이를 감지하는 데 도움이 될 것이다.

기본적으로 현대 AI 시스템은 이해할 수 없는 블랙박스이다.

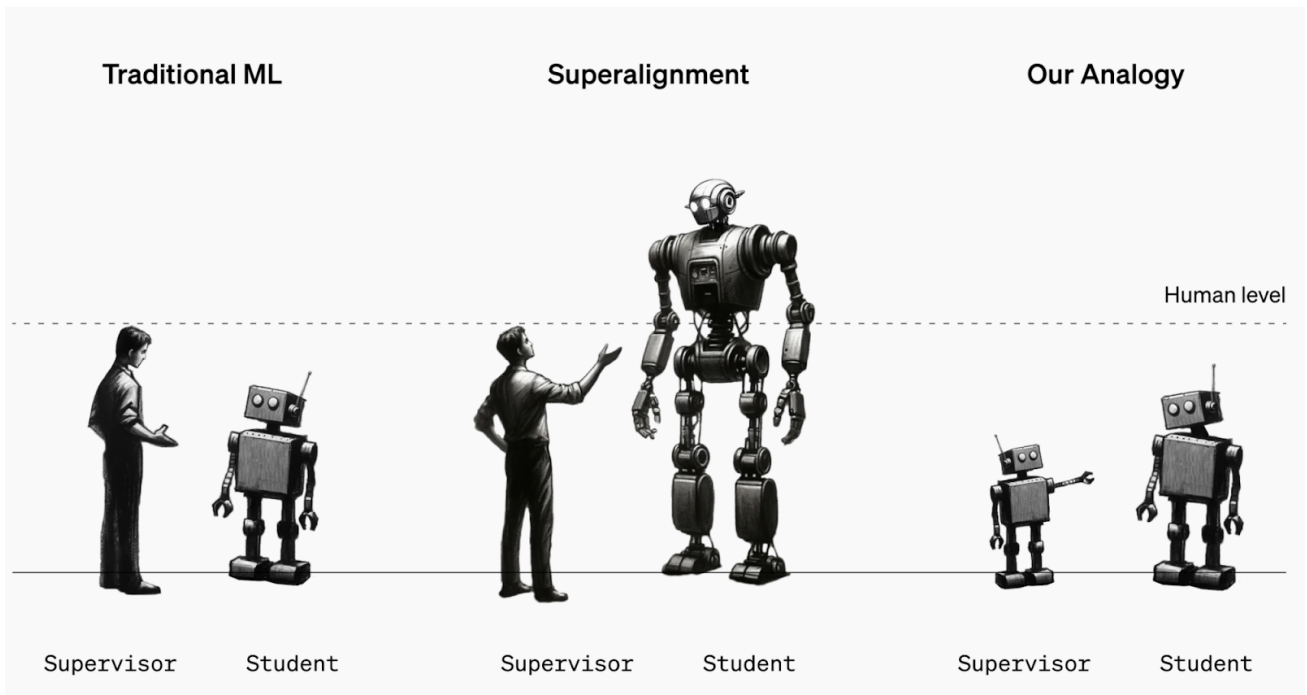


그림 36: 초정렬 연구를 위한 간단한 비유: 초인적 모델을 인간이 감독하는 대신, 작은 모델이 큰 모델을 감독하는 것을 연구할 수 있다. 예를 들어, GPT-4를 GPT-2 감독만으로 정렬할 수 있을까? 그것이 GPT-4가 “GPT-2가 의미한 바”를 적절히 일반화하는 결과를 낳을까? 약한 일반화에서 강한 일반화로.

그럼에도 불구하고 우리는 놀라운 “디지털 신경과학”을 할 수 있을 것 같다—결국 모델 내부에 완벽히 접근할 수 있기 때문이다.

여기에는 몇 가지 접근법이 있는데, “가장 야심차고 ‘멋지지만’ 매우 어려운 것”부터 “더 해킹하기 쉽고 어찌면 작동할 수도 있는 것”까지 다양하다.

기계적 해석 가능성. 대형 신경망을 완전히 역설계하려 시도한다—말하자면 불가해한 행렬을 완전히 분리해내는 것이다.

Anthropic의 Chris Olah 팀이 이 분야에서 많은 선구적 작업을 해왔으며, 아주 작은 모델에서 단순한 메커니즘을 이해하는 것부터 시작했다. 최근 이 분야에서 매우 흥미로운 진전이 이루어지고 있으며, 나는 이 분야 전반의 활동 수준에 매우 기대하고 있다.⁹¹

하지만 나는 초인적 AI 시스템을 완전히 역설계하는 것이 “인간 뇌를 완전히 역설계하는 것”과 유사하게 불가능한 문제일 것이라 우려한다—이 작업을 주로 “AI 안전을 위한 야심찬 달 탐사 프로젝트”로 분류하고, “그냥 대충 해치우는 기본 계획”으로 보지는 않는다.

(Neel Nanda의 기계적 해석 가능성에 관한 200가지 미해결 문제도 이 연구 유형의 감을 준다.)

더욱이, 최근 기계적 해석 가능성 분야에서 희소 오토인코더를 이용해 모델의 “특징을 분리하는” 진전이 있었지만, 이것만으로는 초인적 모델을 다루는 문제를 해결하지 못한다. 예를 들어, 모델이 당신이 이해하지 못하는 초

91. 최근 메커니즘 해석 연구에서 희소 오토인코더로 모델의 ‘특징 분리’를 일부 진전시켰지만, 이것만으로 초인적 모델 처리 문제를 해결하지는 못합니다. 모델은 당신이 이해하지 못하는 초인적 개념으로 ‘생각’할 수 있으며, 원하는 특징이 무엇인지 알기도 어렵습니다. 진실 특징과 보상자 기대 특징, 인간이 생각할 법한 특징 등 여러 특징이 혼재할 수 있습니다. 희소 오토인코더는 도움 되는 도구이긴 하나 ‘일반화 과학’ 같은 문제 해결에 도움이 되도록 결국 활용되어야 합니다.

인적 개념으로 “생각”할 수도 있다. 게다가 어떤 특징이 원하는 것인지 어떻게 알 수 있는가? 여전히 진실 레이블이 없다. 예를 들어, 당신에게 “진실 특징”처럼 보이는 다양한 특징이 많을 수 있는데, 그중 하나는 “모델이 실제로 아는 것”이고, 다른 것들은 “xyz 인간이 어떻게 생각할지” 또는 “인간 평가자가 내가 생각하기를 원하는 것” 등일 수 있다.

희소 오토인코더만으로는 충분하지 않지만, 그것들은 도구가 될 것이다—엄청나게 유용한 도구!—궁극적으로는 일반화 과학과 같은 분야에 도움이 되어야 한다.

“탐다운” 해석 가능성. 기계적 해석 가능성이 신경망을 “바텀업” 방식으로 역설계하려 한다면, 다른 연구들은 보다 목표 지향적인 “탐다운” 접근법을 취해, 정보가 어떻게 처리되는지 완전히 이해하지 않고도 모델 내에서 정보를 찾아내려 한다.

예를 들어, AI 시스템이 거짓말할 때 “활성화되는” 신경망 부분을 식별하여 “AI 거짓말 탐지기”를 만들려고 할 수 있다. 이는 훨씬 더 다루기 쉬울 수 있다(비록 더 강력한 보장을 제공하지는 않더라도).

지난 몇 년간 이 분야에서 흥미로운 연구가 급증했다. CCS는 비지도 데이터만으로 모델 내에서 “진실 방향”을 식별할 수 있다.⁹² ROME은 에펠탑이 파리에 있다는 사실을 아는 모델 부분을 찾아내고, 그 지식을 직접 편집해 에펠탑을 로마에 위치시키기도 한다. 표현 공학과 추론 시 개입은 탐다운 기법을 사용해 거짓말과 환각을 탐지하고, 탈옥(jailbreaking), 권력 추구, 공정성, 진실성 등 모델 행동을 정밀하게 제어하는 것을 보여준다. 모델 내부 정보조차 필요하지 않은 거짓말 탐지에 관한 창의적인 연구도 있다.

나는 탐다운 해석 가능성 기법이 강력한 도구가 될 것이라고 점점 더 낙관한다—즉, “AI 거짓말 탐지기” 같은 것을 만들 수 있을 것이다⁹³—그리고 신경망을 이해하는 데 근본적인 돌파구가 필요하지 않을 것이다.

체인 오브 생각(chain-of-thought, CoT) 해석 가능성. 앞서 언급했듯이, 나는 “생각을 소리 내어 말하는” 체인 오브 생각 시스템으로 AGI를 부트스트랩하는 것이 꽤 그럴듯하다고 본다. 비록 이것이 극한에서는 가장 효율적인 알고리즘이 아닐지라도(초지능이 여전히 영어 체인 오브 생각을 사용하기보다는 어떤 형태의 순환 내부 상태를 통해 사고할 것이라면 매우 놀랄 것이다), 우리는 최초의 AGI에게 이 특성으로부터 여전히 이익을 얻을 수 있다. 이는 해석 가능성에 엄청난 도움이 될 것이다: AGI의 “내부 독백”에 접근할 수 있기 때문이다! 이는 심각한 정렬 실패를 비교적 쉽게 탐지할 수 있게 해줄 것이다.

그러나 이것에 의존하려면 해야 할 일이 많다. CoT가 계속해서 읽기 쉽게 유지되도록 어떻게 보장할 것인가? (계속)

모델을 RL로 훈련시키는 방식에 따라 이해 가능한 영어에서 이해 불가능한 모델 언어로 단순히 변질될 수 있습니다—예를 들어, 이를 가독성 있게 유지하기 위해 간단한 제약을 추가할 수 있을까요? CoT(사고 과정)가 충실한지, 즉 실제로 모델이 생각하는 바를 반영하는지 어떻게 보장할 수 있을까요? (예를 들어, 특정 상황에서 모델이 실제 내부 추론과는 다르게 사후에 CoT에서 추론을 꾸며내는 경우가 있다는 연구가 있습니다.)

92. 본질적으로, 진실의 일관성 속성만 필요하며, 진실/거짓의 강력한 지상 진리(ground truth) 라벨은 필요 없는데, 우리는 초인적 시스템에 대해 후자를 갖지 못합니다.

93. 여전히 많은 기법들이 초인적 모델에서 확장 가능성에 대해 걱정됩니다—대체로 지상 진리 라벨, 즉 모델보다 더 똑똑한 감독자나 우호적 일반화에 의존하는 경향이 있습니다.

제가 생각하기에, 가독성과 충실성에 대한 간단한 측정과 이를 더 오래 유지하기 위한 간단한 해킹이 꽤 효과적인 일 수 있습니다. 물론 어떤 환경에서는 통하지 않을 수 있고 다소 단순한 해킹이지만, 너무나 손쉬운 과실과도 같아서 이 방향은 제 생각에 심각하게 과소평가되고 있습니다.

적대적 테스트 및 측정. 과정 중에, 우리는 매 단계에서 시스템의 정렬(alignment)을 철저히 스트레스 테스트 하는 것이 매우 중요할 것입니다—우리의 목표는 야외에서 실패 모드를 만나기 전에 실험실에서 모든 실패 모드를 경험하는 것입니다. 이를 위해 자동화된 레드팀 기법을 상당히 발전시켜야 합니다. 예를 들어, 모델에 백도어나 비정렬 상태를 의도적으로 심어놓으면, 우리의 안전 훈련이 이를 발견하고 제거할 수 있을까요? (초기 연구는 ‘잠복 요원’이 안전 훈련을 통과할 수 있음을 시사합니다.)

더 일반적으로, 정렬 상태를 잘 측정하는 것이 중요할 것입니다. 모델이 비정렬될 가능성이 있는 힘을 가지고 있나요? 예를 들어, 장기 목표를 가지고 있나요, 그리고 어떤 종류의 동기를 학습하고 있나요? 그리고 명확한 ‘금지선’은 무엇인가요? 예를 들어, 매우 직관적인 경계는 “모델의 추론(사고 과정)은 항상 가독성과 충실성을 유지해야 한다”일 수 있습니다. (에릭 슈미트가 말하길, AI 에이전트가 우리가 이해할 수 없는 언어로 서로 대화할 수 있게 되는 순간, 우리는 컴퓨터를 꺼야 한다고 합니다.) 또 다른 경계는 모델이 완전히 정직한지 여부를 더 잘 측정하는 방법을 개발하는 것일 수 있습니다.

정렬 측정 과학은 아직 초기 단계에 있으며, 이를 개선하는 것은 지능 폭발 시기에 위험에 대한 올바른 균형을 맞추는 데 매우 중요할 것입니다. 과학을 수행하는 것...

우리가 정렬(alignment)을 측정할 수 있게 해주고 “다음 규모 차수(OOM)로 초인간 영역에 진입하는 것이 안전하다는 것을 확신하기 위해 어떤 증거가 충분한가?”에 대한 이해를 제공하는 것은 오늘날 정렬 연구에서 가장 우선순위가 높은 작업 중 하나이다(“다소 초인간적” 시스템으로 RLHF를 확장하려는 작업을 넘어서). 또한 초정렬(superalignment) 연구 방향에 대한 이 글도 참고하라. — 초정렬 패스트 그랜트 제안 요청서.

정렬 연구의 자동화

궁극적으로 우리는 정렬 연구를 자동화해야 할 것이다. 진정한 초지능에 대한 정렬 문제를 직접 해결하는 것은 불가능에 가깝다; 것처럼 거대한 지능 격차를 다루는 것은 매우 도전적이다. 게다가 지능 폭발이 끝날 무렵—1억 명의 자동화된 AI 연구자들이 10년간의 ML 진보를 맹렬히 추진한 후—현재 시스템과 비교해 훨씬 더 이질적인 아키텍처와 알고리즘을 가진 시스템들이 등장할 것으로 예상된다(예를 들어, CoT의 가독성, 일반화 특성, 또는 훈련에 의해 유발되는 정렬 실패의 심각성 등에서 덜 우호적인 특성을 가질 수 있다).

하지만 이 문제를 우리 혼자서만 해결할 필요는 없다. 다소 초인간적 시스템을 충분히 정렬하여 신뢰할 수 있게 된다면, 우리는 엄청난 위치에 서게 될 것이다: 최고의 AI 연구자보다 더 똑똑한 수백만 명의 자동화된 AI 연구자들을 마음껏 활용할 수 있게 된다. 이 자동화된 연구자 군단을 적절히 활용하여 더욱 초인간적인 시스템에 대한 정렬 문제를 해결하는 것이 결정적일 것이다.

(참고로, 이는 오용 등 AI 위험 전반에 더 일반적으로 적용된다. 모든 경우에 AI 안전을 위한 최선의 경로—아마도 유일한 경로—는 초기 범용 인공지능(AGI)을 안전에 적절히 활용하는 것이다; 예를 들어, 일부 AGI를 외국 행위자가 모델 가중치를 빼내는 것을 방어하는 자동화 연구에 투입하고, 다른 일부는 최악의 생물학적 공격에 대비한 방어 강화에 투입하는 식이다.)

지능 폭발 동안 자동화된 정렬을 제대로 수행하는 것은 극히 중요한 사안이 될 것이다: 우리는 몇 달 만에 수년간의 AI 발전을 겪게 될 것이며, 거의

인간이 올바른 결정을 내리는 데 필요한 시간이 줄어들면서, 우리는 정렬 실패가 치명적일 수 있는 영역에 진입하기 시작할 것입니다. 연구소들은 필요하다면 지능 폭발 동안 자동화된 정렬 연구(자동화된 능력 연구 대비)에 컴퓨팅 자원의 상당 부분을 투입할 의지가 있어야 합니다. 우리는 자동화된 정렬 연구 결과를 신뢰할 수 있게 해주는 강력한 보장과, 오늘날보다 훨씬 더 나은 비정렬 측정 방법이 필요합니다. 그래야 우리가 여전히 안전한지 알 수 있습니다. 우리가 상승하고자 하는 각 규모 차수(OOM)마다, 다음 규모 차수에 대한 정렬 접근법에 대해 극도로 높은 신뢰를 가져야 할 것입니다.

무엇보다도, 이를 올바르게 수행하려면 극도의 역량, 진지함, 그리고 어려운 선택을 감수하려는 의지가 필요합니다. 진정한 초지능에 가까워질수록 위험은 매우 현실적이 되며, 실패—악성 AI—는 치명적일 수 있습니다. 증거가 모호한 순간들이 있을 것이며, 지표들은 겉보기에는 괜찮아 보이지만, 임박한 위험을 암시하는 몇 가지 경고 신호가 있을 것입니다. 그리고 우리는 대체로 무슨 일이 벌어지고 있는지 전혀 알지 못할 것입니다. 이 상황을 책임감 있게 헤쳐 나가려면 매우 큰 비용이 드는 선택을 해야 합니다. 결국, 차세대 초지능에 대한 승인 결정은 군사 작전 개시 결정만큼 진지하게 다뤄져야 할 것입니다.

초방어

“정렬을 올바르게 하는 것”은 지능 폭발 동안 여러 방어층 중 첫 번째에 불과해야 합니다. 정렬은 어려울 것이며, 그 과정에서 실패가 있을 것입니다. 가능하다면, 우리는 정렬이 실패할 수 있는 상황에 있어야 하지만—그 실패가 치명적이지는 않아야 합니다. 이는 다음을 의미할 수 있습니다:

- 보안. 에어갭 클러스터는 초지능이 스스로 정보를 빼내어 현실 세계에 피해를 주려는 시도에 대한 첫 번째 방어층입니다. 그리고 그것은 시작에 불과합니다; 우리는 모델의 자체 정보 유출에 대해 훨씬 더 극단적인 보안이 필요할 것입니다⁹⁴. 하드웨어 암호화부터 다중 키 승인⁹⁵에 이르기까지 전반적으로 강화해야 합니다. 그리고 우리의 범용 인공지능(AGI)이 자동화된 방식으로 정보를 빼낼 수 있게 되면, 인간이 AI에 속거나 설득되어 정보를 빼내는 것을 방지하는 것도 필요합니다.

연구개발(R&D) 측면에서도, 보안을 더욱 강화하는 데 이들을 투입해야 합니다.

- 모니터링. 우리의 AI 시스템이 수상하거나 악의적인 행동을 하거나, 혹은 내부 직원이 무단 활동에 AI를 사용하려 할 경우 이를 포착할 수 있어야 합니다. 그러나 이는 간단한 일이 아닙니다. 수백만 개의 범용 인공지능(AGI) 인스턴스가 클러스터에서 복잡한 코드를 실행하고 있을 것이기 때문입니다. 모델 출력, 사고 과정, 모델이 작성한 코드, 서버/네트워크 활동, 연구원 및 기타 인간과의 상호작용 등 모델 자체를 활용하는 고급 모니터링 시스템이 필요합니다. 본질적으로 일부 AI 시스템을 이용해 다른 AI 시스템을 감시하는 셈입니다. 또한, 덜 강력하지만 신뢰할 수 있는 모델을 사용해 더 강력하지만 덜 신뢰할 수 있는 모델의 전복을 방지하는 정교한 ‘통제’ 프로토콜을 개발하고 적용해야 합니다. 예를 들어, 활성 모델 인스턴스에 할당하는 컴퓨팅 자원과 동일한 양을 모니터링 모델 인스턴스에 할당해야 할 수도 있습니다.

94. 모델이 자신의 가중치를 훔쳐 원래 데이터센터 밖에서 자신을 복제하는 행위.

95. AI에 속거나 설득되어 정보를 빼내는 인간을 방어하는 것.

- 목표 능력 제한. 가능한 한, 실패 시 피해를 줄일 수 있도록 모델의 능력을 목표 지향적으로 제한해야 합니다. 대표적인 예는 생물학 및 화학과 관련된 모든 내용을 모델 학습에서 제거하거나 ‘언러닝(unlearning)’ 기법을 사용하는 것입니다. 이는 생물학 및 화학 무기를 통한 치명적인 현실 세계 피해 경로를 차단하는 데 도움이 됩니다.

- 목표 학습 방법 제한. 일부 학습 방법은 본질적으로 더 위험하며, 심각한 비정렬을 초래할 가능성이 더 높을 수 있습니다. 예를 들어, 모방 학습은 비교적 안전해 보입니다(위험한 장기 내부 목표를 가진 모델이 만들어질 가능성이 낮음). 반면, 장기 결과 기반 강화학습(RL)은 가능한 한 피해야 합니다(모델이 바람직하지 않은 장기 목표를 학습할 가능성이 훨씬 높음). 또한, 해석 가능성 방법과 모니터링 체계에 반하는 학습도 피해야 합니다(모델이 이를 회피하도록 학습되는 것을 방지). 추가로 바람직한 제한이 있을 수 있는데, 예를 들어 사고 과정이 명료하고 충실하게 유지되도록 하는 것 등이 있습니다.

가능한 한, 우리는 이러한 제약 조건을 미리 정의하고, 지능 폭발이 진행되는 동안 최대한 오래 유지하며, 절대적으로 필요할 때만 이를 해제해야 합니다.

- 그리고 여기에는 훨씬 더 많은 가능성이 있을 것입니다.

이것들이 완벽할까요? 전혀 그렇지 않습니다. 진정한 초지능은 대부분의 보안 체계를 우회할 수 있을 가능성이 큼니다. 그럼에도 불구하고, 이러한 조치는 우리에게 훨씬 더 많은 오류 여유를 제공합니다—그리고 우리는 가능한 모든 여유를 필요로 할 것입니다. 우리는 그 여유를 활용해 정렬 기술에 대해 매우 높은 신뢰를 가질 수 있는 위치에 도달하고, 신뢰에 따라 “초방어” 조치(예: 초지능을 비격리 환경에 배치하는 것)를 점진적으로 완화하고자 할 것입니다.

이러한 AI 시스템을 덜 통제된 환경, 예를 들어 군사적 용도로 배치할 때 상황은 다시 불안정해질 것입니다. 상황이 우리를 빠르게 그렇게 하도록 강요할 가능성이 크지만, 우리는 항상 가능한 한 많은 오류 여유를 확보하려고 노력해야 합니다—예를 들어, 초지능을 군사 목적으로 “현장”에 직접 배치하는 대신, 더 격리된 환경에서 연구 개발에 활용하고, 그들이 발명한 특정 기술(예: 우리가 더 신뢰할 수 있는 제한된 자율 무기 시스템)만 배치하는 방식입니다.

내가 낙관적인 이유, 그리고 두려운 이유

나는 초정렬 문제의 기술적 해결 가능성에 대해 매우 낙관적입니다. 이 분야에는 곳곳에 손쉽게 얻을 수 있는 성과들이 많아 보입니다. 더 넓게 보면, 딥러닝의 경험적 현실은 10년 전 일부가 예상했던 것보다 우리에게 더 유리하게 전개되고 있습니다. 예를 들어, 딥러닝은 많은 상황에서 놀라울 정도로 온건하게 일반화합니다: 종종 우리가 의도한 대로 “그 일을 수행”할 뿐, 어떤 난해하고 악의적인 행동을 학습하지 않습니다.⁹⁶ 게다가,

모델 내부를 완전히 이해하는 것은 어려울 것이지만, 적어도 초기 범용 인공지능(AGI)에 대해서는 해석 가능성에 대한 꽤 좋은 가능성이 있습니다—우리는 사고의 연쇄(chain of thought)를 통해 투명하게 추론하도록 만들 수 있으며,

표현 공학과 같은 임시방편 기법들은 “거짓말 탐지기”나 이와 유사한 역할로 놀랍도록 잘 작동한다.

96. 물론, 이것이 현재 모델과 아마도 인간 수준 시스템에 적용될 수 있지만, 미래의 훨씬 초인적인 모델에 대해 현재 모델에서 얻은 증거를 외삽하려 할 때는 주의해야 합니다.

나는 “어느 정도 초인적인” 시스템을 정렬하는 “기본 계획”이 대체로 성공할 가능성이 꽤 합리적이라고 생각한다.⁹⁷

물론, 추상적으로 “기본 계획”에 대해 말하는 것과 그 계획을 실행할 책임이 있는 팀이 당신과 20명의 동료라는 것은 전혀 다른 문제다(훨씬 더 스트레스가 크다!).⁹⁸

“어떤 회사가 거대한 방을 채우고, 한 도시만큼의 전기를 소비하며, 최근에 사람처럼 대화하는 놀라운 능력을 얻은 AI를 만들고 있다. 이 AI는 어떤 주제에 대해서도 에세이나 시를 쓸 수 있고, 대학 수준의 시험을 통과할 수 있다. 매일 새로운 능력을 얻고 있는데, AI를 관리하는 엔지니어들조차 공개적으로 말할 수 없는 것들이다. 하지만 그 엔지니어들은 회사 식당에 앉아 자신들이 만드는 것의 의미에 대해 토론한다. 다음 주에는 무엇을 배우게 될까? 어떤 직업들이 쓸모없게 될까? 그들이 속도를 늦추거나 멈춰야 할까, 용의 꼬리를 건드리지 않도록? 하지만 그렇다면 누군가, 아마도 덜 양심적인 누군가가 먼저 용을 깨우는 게 아닐까? 세상에 더 많이 알려야 할 윤리적 의무가 있을까? 아니면 덜 알려야 할 의무가 있을까? 나는—너는—그 회사에서 1년을 보내고 있다. 내 일—네 일은 AI와 그 후계자들이 대혼란을 일으키지 못하도록 수학적 이론을 개발하는 것이다. 여기서 ‘대혼란’은 선전 선동과 학문적 부정행위 가속화부터 생물 테러 조연 제공, 그리고 그렇다, 세상을 파괴하는 것까지 무엇이든 의미할 수 있다.”

이것은 계획의 첫 부분일 뿐이다—내가 밤에 잠 못 이루게 하는 진짜 문제는 지능 폭발이다. 첫 번째 범용 인공지능, 즉 어느 정도 초인적인 시스템을 정렬하는 것은 한 가지 문제다.

훨씬 초인적이고 이질적인 초지능은 완전히 다른 문제이며, 그것은 무서운 문제다.

지능 폭발은 제품 출시라기보다는 전쟁을 치르는 것과 더 비슷할 것이다. 우리는 초방어(superdefense)나 에어갭 클러스터 같은 것을 준비하고 있지 않으며, 모델이 스스로 데이터를 빼돌려도 알아차리지 못할 수도 있다. 우리는 이 엄청나게 위험한 결정들을 내릴 합리적인 지휘 체계를 갖추고 있지 않으며, 초지능에 대해 매우 높은 신뢰도를 요구하거나, 다음 훈련 실행을 시작하기 전에 안전을 확보하기 위해 추가 시간을 들이거나, 정렬 연구에 컴퓨팅 자원의 대부분을 할당하는 어려운 결정을 내릴 준비가 되어 있지 않다. 위험을 인지하고 충돌을 피하기보다는 그냥 부딪힐 것이다. 현재 어떤 연구소도 안전을 확보하기 위해 비용이 많이 드는 절충을 할 의지가 거의 없음을 보여주고 있다(안전 위원회는 많지만, 그건 별 의미가 없다). 기본적으로 우리는 아마도 지능 폭발에 우연히 휘말리게 될 것이고, 사람들이 우리가 무엇에 빠져들었는지 깨닫기도 전에 몇 차례 규모 차수(OOM)의 도약을 겪게 될 것이다.

우리는 운에 너무 많이 기대고 있다.

IIId. 자유 진영이 반드시 승리해야 한다

초지능은 결정적인 경제적, 군사적 우위를 제공할 것이다. 중국은 아직 완전히 게임에서 벗어나지 않았다. 범용 인공지능(AGI) 경쟁에서 자유 진영의 생존 자체가 위태로울 것이다. 우리는 권위주의 세력에 대한 우위를 유지할 수 있을까? 그리고 그 과정에서 자기 파괴를 피할 수 있을까?

인류 역사는 전쟁의 역사다. 잠시 동안의 불안정한 휴전기를 제외하면, 세계에는 평화가 없었고; 역사가 시작되기 전부터 살인적인 갈등은 보편적이고 끝이 없었다. ... 오렌지만 한 크기의 폭탄이 건물 한 블록 전체를 파괴하

는 비밀스러운 힘을 가질 수 있지 않을까 — 아니, 천 톤의 코다이트 힘을 집중시켜 한 번에 마을을 폭파할 수 있지 않을까? 윈스턴 처칠 “우리 모두 자살할 것인가?” (1924)

초지능은 인류가 개발한 가장 강력한 기술이자 가장 강력한 무기가 될 것이다. 이는 핵무기와 견줄 만한 결정적인 군사적 우위를 제공할 수 있다. 권위주의자들은 초지능을 세계 정복과 내부 완전 통제에 사용할 수 있다. 불량 국가들은 이를 이용해 전멸 위협을 가할 수도 있다. 많은 이들이 중국공산당(CCP)을 제외한 것으로 보지만, CCP가 AGI의 중요성을 깨닫는 순간 경쟁력을 갖출 명확한 길이 있다(적어도 우리가 미국 AI 연구소의 보안을 대폭 강화하지 않는 한).

안전 측면에서도 매달의 선도 기간이 중요하다. 우리는...

가장 큰 위험은 우리가 치열한 경쟁에 갇혀 있을 때 발생한다. 민주주의 동맹국들과 권위주의 경쟁자들이 이미 불안정한 지능 폭발을 숨 가쁘게 지나가며 서로가 먼저 초지능을 획득할까 두려워 조심성을 완전히 버려야 하는 상황이다. 민주주의 동맹국들이 건강한 우위를 유지할 때만이 초지능 출현이라는 극도로 불안정하고 위험한 시기를 헤쳐 나갈 수 있는 오차 범위를 확보할 수 있다. 그리고 미국의 리더십만이 초지능이 초래할 자기 파괴 위험을 막기 위한 비확산 체제를 개발하는 현실적인 길이다.

우리 세대는 평화와 자유 속에 살고 있다는 사실을 너무 쉽게 당연하게 여긴다. 그리고 SF에서 AGI 시대를 예고 하는 이들은 너무 자주 방 안의 코끼리, 즉 초지능이 국가안보 문제라는 점을 무시한다. 미국은 반드시 승리해야 한다.

누가 초지능을 선도하느냐가 결정적인 군사적 우위를 가져올 것이다

초지능은 단순한 기술이 아니다—극초음속 미사일, 스텔스 등과 같은 기술로, 미국과 자유 민주주의 국가들의 리더십이 매우 바람직하지만 절대적으로 필수적인 것은 아니다. 미국이 한두 가지 기술에서 뒤쳐져도 군사력 균형은 유지될 수 있다; 이러한 기술들은 매우 중요하지만 다른 분야의 이점으로 상쇄될 수 있다.

초지능의 출현은 원자력 시대 이후로 본 적 없는 상황을 초래할 것이다: 초지능을 가진 자가 가지지 못한 자를 완전히 지배하게 된다.

나는 이전에 초지능의 막대한 힘에 대해 논의한 바 있다. 초지능은 가장 똑똑한 인간 과학자보다 훨씬 뛰어난 수십억 명의 자동화된 과학자, 엔지니어, 기술자들이 밤낮없이 새로운 기술을 미친 듯이 발명하는 것을 의미한다. 과학 및 기술 발전의 가속은 엄청날 것이다. 초지능이 군사 기술 연구개발에 적용되면 우리는 군사 기술 발전의 수십 년을 빠르게 통과할 수 있다.

걸프 전쟁, 혹은 수십 년간의 기술 우위가 군사력에 의미하는 바

걸프 전쟁은 군사 기술에서 20~30년의 우위가 얼마나 결정적일 수 있는지를 보여주는 유용한 사례를 제공합니다. 당시 이라크는 세계에서 네 번째로 큰 군대를 보유하고 있었습니다. 병력, 전차, 포병 등 숫자 면에서 미 주도 연합군은 간신히 맞먹거나 오히려 이라크군에 밀렸으며, 이라크군은 방어선을 충분히 구축한 상태였습니다(이런 상황에서는 군사 병력에서 3: 또는 5:의 우위가 있어야 교란이 가능합니다).

하지만 미 주도 연합군은 단 100시간의 지상전만에 이라크 군대를 완전히 격파했습니다. 연합군 사상자는 단 292명에 불과한 반면, 이라크 사상자는 2만~5만 명에 달했고 수십만 명이 부상하거나 포로가 되었습니다. 연합군은 전차 31대만 잃었으나, 이라크 전차는 3,000대 이상 파괴되었습니다.

기술 차이는 신적이거나 이해 불가능한 수준은 아니었지만, 완전히 결정적이었습니다. 유도 및 스마트 탄약, 초기 스텔스 기술, 더 나은 센서, 더 먼 거리에서 야간과 먼지 폭풍 속을 볼 수 있는 전차 조준경, 더 우수한 전투기, 정찰 우위 등이 그것입니다.

(더 최근 사례로는 이란이 이스라엘에 300발의 미사일을 대규모로 발사했으나, 이스라엘과 미국, 동맹국의 우수한 미사일 방어 시스템에 의해 “99%”가 요격된 것을 떠올려 보십시오.)

범용 인공지능(AGI)에서 1~3년의 우위는 걸프 전쟁에서 미 연합군이 이라크에 가졌던 것만큼 결정적인 군사적 우위를 의미할 수 있습니다. 군사력 균형의 완전한 재편이 걸려 있을 것입니다.

만약 우리가 20세기의 군사 기술 발전을 10년도 채 안 되는 기간에 겪었다고 상상해 보십시오. 우리는 몇 년 만에 말과 소총, 참호에서 현대 전차 군대로, 그리고 그로부터 몇 년 후에는 초음속 전투기 함대와 핵무기, 대륙간 탄도미사일(ICBM)로 발전했을 것입니다.

그것; 적이 당신이 거기에 있다는 사실조차 알기도 전에 적을 무력화할 수 있는 은밀함과 정밀함, 그로부터 몇 년 후에.

이것이 바로 초지능의 출현과 함께 우리가 직면할 상황이다: 한 세기의 군사 기술 발전이 10년도 채 안 되는 기간으로 압축된다. 우리는 적의 군사력을 마비시킬 수 있는 초인적 해킹, 로봇 군대와 자율 드론 떼를 보게 될 것이다. 하지만 더 중요한 것은 아직 상상조차 할 수 없는 완전히 새로운 패러다임과 파괴력이 천 배 이상 증가한 새로운 대량살상무기(WMD)의 발명(그리고 불침투 미사일 방어 같은 새로운 WMD 방어 수단도 함께 등장하여 억지력 균형을 빠르고 반복적으로 뒤흔든다).

그리고 이것은 단순한 기술 진보에 그치지 않는다. 로봇공학이 해결되면 노동이 완전 자동화되어 더 광범위한 산업 및 경제 폭발도 가능해진다. 연간 10%대의 성장률도 충분히 가능하며, 최대 10년 내에 선도하는 국가의 GDP는 뒤쳐진 국가들을 압도할 것이다. 급속히 증가하는 로봇 공장들은 단지 기술적 우위뿐 아니라 순수 물질적 생산 능력에서도 지배력을 의미한다. 수백만 개의 미사일 요격기, 수십억 대의 드론 등을 생각해 보라.

물론 과학의 한계와 여러 마찰 요인들이 속도를 늦출 수도 있다. 하지만 결정적인 군사 우위를 위해 신적인 발전은 필요 없다. 그리고 10억 명의 초지능 과학자들이 할 수 있는 일은 많다. 초지능 이전의 군대는 몇 년 내에 압도적으로 뒤쳐질 것이 분명해 보인다.

군사적 우위는 핵 억지력에 대해서도 결정적일 것이다

더 명확히 말하자면: 초지능이 부여하는 우위는 적의 핵 억지력을 선제적으로 제거할 만큼 결정적일 가능성이 높다. 향상된 센서 네트워크와 분석 능력은 가장 은밀한 현재의 핵 잠수함도 찾아낼 수 있을 것이다(이동식 미사일 발사대도 마찬가지다). 수백만 또는 수십억 마우스 크기의

자율 드론은 스텔스 기술의 발전과 함께 적 후방에 침투하여 은밀히 적의 핵무기 부대를 탐지, 파괴, 그리고 지휘부를 제거할 수 있다. 향상된 센서, 표적 지정 기술 등은 미사일 방어 능력을 극적으로 향상시킬 수 있다(예를 들어 위의 이란 대 이스라엘 사례와 유사하게); 게다가 산업적 폭발이 일어난다면, 로봇 공장들이 상대 미사일 하나당 수천 대의 요격 미사일을 생산할 수 있다. 그리고 이 모든 것은 완전히 새로운 과학 및 기술 패러다임(예: 원격으로 모든 핵무기를 비활성화하는 기술)을 고려하지 않은 상태다.

이런 상황에서는 경쟁이 성립하지 않는다. 여기서 ‘경쟁이 성립하지 않는다’는 것은 단순히 “서로를 상호 파괴할 수 있다”는 핵무기적 의미뿐만 아니라, 상당한 인명 피해 없이 경쟁자의 군사력을 완전히 파괴할 수 있다는 의미다. 초지능에서 몇 년의 선도는 완전한 지배를 의미할 것이다.

만약 지능 폭발이 급격히 일어난다면, 몇 개월의 선도만으로도 결정적일 수 있다: 몇 개월은 대략 인간 수준의 AI 시스템과 상당히 초인적인 AI 시스템 간의 차이를 의미할 수 있다. 아마도 초기 초지능만 보유하고 있어도, 광범위하게 배치되기 전이라도 결정적인 우위가 될 수 있다. 예를 들어, 초인적 해킹 능력으로 초지능 이전의 군대를 마비시키거나, 상대 지도자, 관리 및 그 가족 모두에게 즉각적인 죽음을 위협하는 제한된 드론 무리, 그리고 AlphaFold 스타일의 시뮬레이션으로 개발된 특정 민족 집단을 겨냥한 고급 생물무기(예: 한쪽을 제외한 누구나 대상이 되거나 단순히 적에게 치료제를 제공하지 않는 경우) 등이 가능하다.

중국은 경쟁력이 있다

많은 사람들이 중국과 범용 인공지능(AGI)에 대해 단일하게 생각하는 것 같다. 칩 수출 통제로 인해 중국은 무력화되었고, 주요 AI 연구소는 미국과 영국에 있으니 크게 걱정할 필요가 없다고 생각한다. 중국의 대형 언어 모델(LLM)은 괜찮다—분명히 대형 모델을 훈련시킬 능력이 있다!—하지만 그들은 기껏해야 ...

미국 2차 연구소 수준.⁹⁹ 그리고 중국 모델들조차도 종종 미국 오픈 소스 릴리스의 단순 모방에 불과하다(예: Yi-34B 아키텍처는 본질적으로 Llama2 아키텍처이며, 코드 몇 줄만 변경되었다¹⁰⁰). 중국의 딥러닝은 예전만큼 중요하지 않다(예: 바이두는 최초의 현대적 스케일링 법칙 논문 중 하나를 발표했음). 중국은 미국보다 AI 논문을 더 많이 발표하지만, 최근 몇 년간 주요 혁신을 주도하지는 못한 것으로 보인다.

하지만 이것은 단지 서막에 불과하다. 만약 중국공산당(CCP)이 범용 인공지능(AGI)에 눈을 뜬다면, CCP가 경쟁하기 위해 엄청난 노력을 기울일 것으로 예상해야 한다. 그리고 중국이 경쟁에 참여할 수 있는 꽤 명확한 경로가 있다고 생각한다: 미국을 능가하는 인프라를 구축하고 알고리즘을 탈취하는 것이다.

1. 컴퓨팅

1a. 칩: 중국은 이제 7nm 칩 제조 능력을 입증한 것으로 보인다. 7nm를 넘어서는 것은 어렵겠지만(EUV 필요), 7nm면 충분하다! 참고로 7nm는 엔비디아 A100이 사용한 공정이다. SMIC 7nm 플랫폼 기반의 화웨이 Ascend 910B는 동급 엔비디아 칩 대비 성능/가격이 약 2~3배 떨어지는 것으로 보인다.¹⁰¹

SMIC의 7nm 생산 수율과 중국의 전반적인 역량 성숙도는 논란이 있으며,¹⁰² 중요한 미해결 질문은 이 7nm 칩을 어느 정도 규모로 생산할 수 있는냐이다.¹⁰³ 그럼에도 불구하고, 몇 년 내에 대규모 생산이 가능할 합리적인 가능성이 있어 보인다.

99. 예를 들어, Yi-Large는 LMSys에서 GPT-4 급 모델로 보이지만, 이는 OpenAI가 GPT-4를 공개한 지 1년 이상 지난 시점입니다.

100. 유사하게, Qwen의 hugging face 코드에서 Mistral을 많이 참조하며, 중국 LLM의 미국 오픈소스 의존이 중국 총리가 우려하는 명백한 문제입니다.

101. 화웨이 Ascend 910 B는 카드당 약 12만 위안, 즉 약 1만 7천 달러로 보입니다. SMIC 7nm 공정에서 생산되며 A100과 비슷한 성능입니다. H100은 A100 대비 약 3배 높은 성능이며, 가격도 다소 높아(2~2.5만 달러 예상) 중국 내 AI GPU 성능 대비 비용 상승은 약 2~3배 수준입니다.

102. 예를 들어, 그들은 여전히 서방산 HBM 메모리를 사용 중인데(이것이 왜 수출 통제 대상이 아닌지는 의문), CXMT는 내년에 HBM 샘플링을 시작할 예정입니다.

AI 칩의 대부분 성능 향상은 AI 용도에 맞게 칩 설계를 개선한 데서 왔으며(중국은 이미 대만 공급망에서 엔비디아 칩 설계를 탈취한 것으로 보인다¹⁰⁴), 7nm와 3nm 또는 2nm 공정 차이, 그리고 제조 초기 단계의 미성숙함은 중국에 비용 부담을 더할 것이다.¹⁰⁵ 하지만 이것이 치명적이라고 보기는 어렵다; 7nm 공정으로도 매우 좋은 AI 칩을 만들 수 있다. 예를 들어, 1000억 달러 이상, 조 단위 달러 규모의 예산을 투입하면 충분한 컴퓨팅을 확보할 수 없을 것이라고 확신하지는 않는다.

화웨이 Ascend 910B는 카드당 약 12만 위안, 즉 약 1만 7천 달러 정도로 추정된다. 이는 SMIC 7nm 노드에서 생산되며, A100과 유사한 성능을 낸다. H100은 A100 대비 약 3배 성능이 좋으며, 가격은 약간 더 비싼 2~2.5만 달러 수준으로, 중국 AI GPU 성능 대비 비용은 약 2~3배 증가한 셈이다.

예를 들어, 여전히 서방산 HBM 메모리를 사용 중인데(이것이 왜 수출 통제 대상이 아닌지는 미스터리), CXMT는 내년에 HBM 샘플링을 시작할 예정이라고 한다.

서방에서 다른 종류 칩을 여전히 수입할 수 있으므로, 7nm 노드를 AI 칩 생산에 전량 투입해 전체 생산량 감소를 보완할 수 있다.

사이버 범죄자들도 엔비디아를 해킹해 주요 GPU 설계 비밀을 입수했다. 또한 최근 구글에서 기소된 중국 국적자가 TPUv⁶ 설계도 탈취한 것으로 알려졌다.

예를 들어, 성능/가격 또는 성능/전력 효율이 2배 정도 떨어질 수 있다. 이는 동일한 데이터센터 성능을 달성하려면 더 많은 전력과 더 많은 칩을 네트워크로 연결해야 하므로 운영이 더 복잡해진다는 의미다.

몇 년 내에 훈련 클러스터.¹⁰⁶

103. 서방에서 다른 유형 칩 수입이 가능하므로, 전체 7nm 공정을 AI 칩에 전부 투입하여 전체 생산량 감소를 상쇄할 수 있습니다.

104. 특히 사이버범죄자들이 엔비디아를 해킹해 주요 GPU 설계 비밀을 탈취했습니다. TPUv6 설계 또한 최근 기소된 구글 내 중국 국적자가 훔친 것에 포함됩니다.

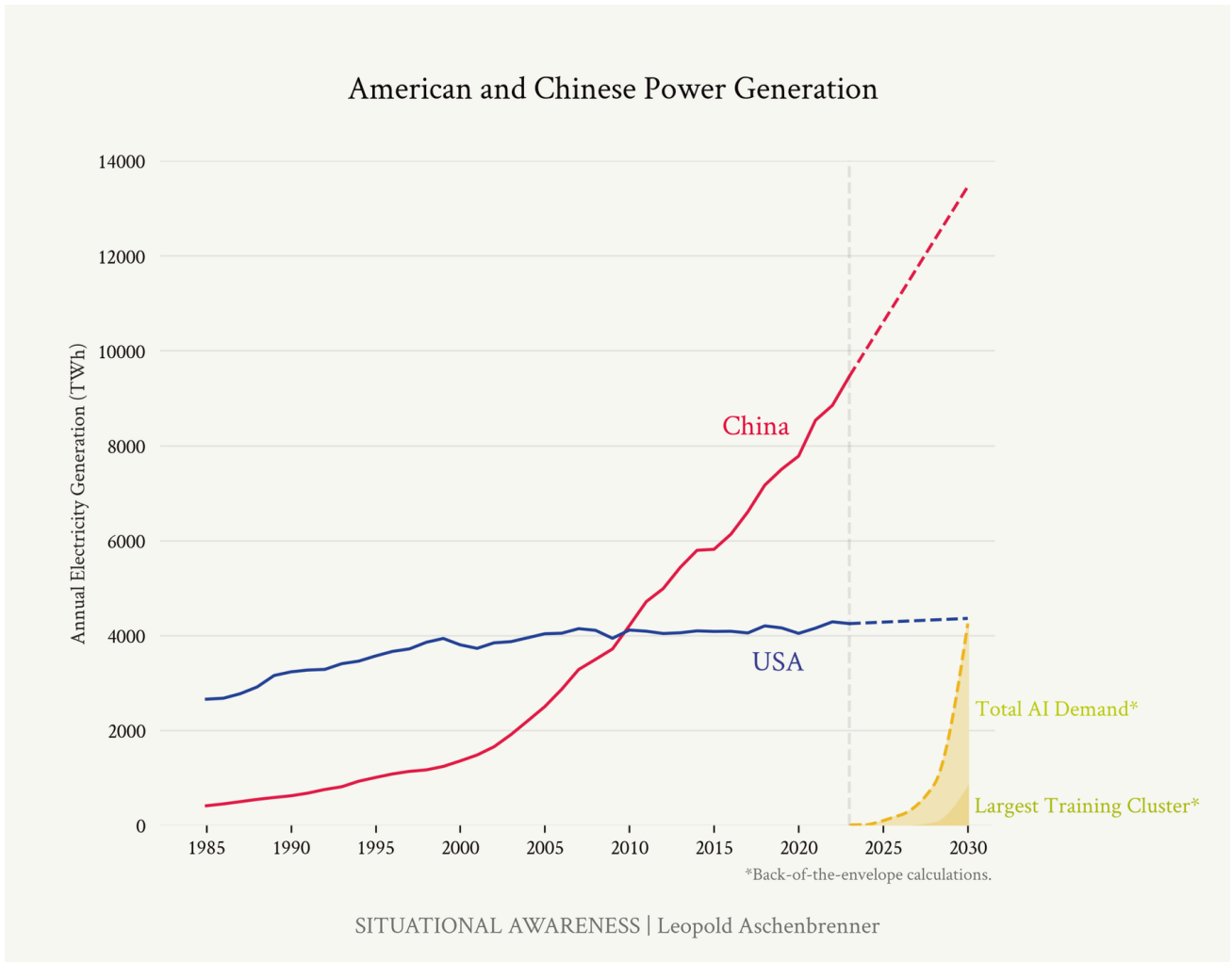


그림 37:2030년 AI 전력 확장은 미국보다 중국이 훨씬 더 실현 가능해 보인다. IIIa의 조 단위 클러스터 경쟁 초기 추정치를 기반으로 함.

1. 알고리즘

OOM 규모 차수를 세는 작업에서 광범위하게 논의했듯이, 컴퓨팅 확장은 이야기의 일부일 뿐이다. 알고리즘 발전이 AI 진보의 적어도 절반 이상을 차지할 가능성이 크다. 우리는 지금 범용 인공지능(AGI)을 위한 핵심 알고리즘 혁신을 개발 중이다(데이터 장벽 때문에 사실상 알고리즘의 EUV와 같다).

105. 예를 들어, 이 칩들이 성능 대비 비용이나 성능 대비 전력 면에서 2배 이상 나쁠 수 있습니다. 결과적으로 데이터센터 성능을 같게 하려면 더 많은 전력과 더 많은 칩을 네트워크로 연결해야 하며, 이는 추가적인 부담입니다.

6. 원 논문에 따르면: “우리는 또한 MATH에서 인간을 평가했으며, 특히 수학을 좋아하지 않는 컴퓨터 과학 박사 과정 학생이 약 40%를 달성한 반면, 세 차례 IMO 금메달리스트는 90%를 달성하여 MATH가 인간에게도 도전적임을 나타냈습니다.”

106. 칩 비용이 3배 증가해도 데이터센터 비용 증가 측면에서는 훨씬 적다는 점에 유의하라. 실제로 논리 반도체 제조 비용은 1b. 미국을 능가하기: 가장 큰 훈련 클러스터에 대한 제약은 칩이 아니라 산업 동원력— Nvidia GPU 비용의 5% 미만이며, 메모리와 CoWoS를 고려해도 Nvidia 가격표의 20% 미만이다. 아마도 조 단위 클러스터에 필요한 100GW 전력일 것이다. 하지만 중국이 미국보다 그리고 GPU 자체도 데이터센터 비용의 50-60%에 불과한 경향이 있다. 더 잘할 수 있는 한 가지가 있다면 바로 건설이다. 따라서 칩 제조 비용이 3배 비싸져도 전체 비용이 3배 증가하는 것과는 거리가 멀다. 지난 10년간 중국은 미국 전체 전력 용량과 맞먹는 신규 전력 용량을 대략 건설했다(미국 용량은 칩 비용이 10배 비싸더라도 중국이 데이터센터 비용을 크게 사실상 정체 상태였다). 미국에서는 이런 사업들이 환경 심사, 허가, 규제로 10년씩 지연된다. 증가시키지 않고도 감당할 수 있을 것 같다. 따라서 중국이 가장 큰 훈련 클러스터에서 단순히 미국을 능가할 가능성은 매우 높아 보인다.

기본적으로, 나는 서구 연구소들이 훨씬 앞서 있을 것으로 예상한다; 그들은 핵심 인재를 많이 보유하고 있으며, 최근 몇 년간 모든 주요 돌파구를 개발해왔다. 이 우위의 규모는 몇 년 내에 10배(심지어 100배) 더 큰 클러스터에 해당할 수 있다; 이는 미국에 상당히 편안한 선두를 제공할 것이다.

그럼에도 불구하고, 현재의 경로대로라면 우리는 이 우위를 완전히 포기할 것이다: 보안 섹션에서 광범위하게 논의했듯이, 현재의 보안 상태는 중국이 미국 연구소에 침투하는 것을 본질적으로 매우 쉽게 만든다. 따라서 연구소를 매우 빨리 잠그지 않는 한, 나는 중국이 AGI의 핵심 알고리즘 요소를 단순히 훔쳐 미국의 역량과 맞먹을 수 있을 것으로 예상한다.¹⁰⁷

(더 나쁜 점은, 우리가 보안을 개선하지 않는다면, 중국이 경쟁할 수 있는 더 명확한 경로가 있다는 것이다. 그들은 심지어 자체 AGI를 훈련할 필요조차 없을 것이다: 그들은 AGI 가중치를 직접 훔칠 수 있다. 자동화된 AI 연구원 한 부류를 훔친 후, 그들은 곧바로 경주에 나설 수 있으며, 자신들의 지능 폭발을 시작할 수 있다. 만약 그들이 미국보다 덜 신중하다면—좋은 신중함과 불합리한 규제 및 지연 모두를 포함하여—그들은 지능 폭발을 더 빠르게 진행하여 우리를 앞지를 수도 있다.)

지금까지 미국 기술 기업들은 중국의 어떤 노력보다 AI와 스케일링에 훨씬 더 큰 베팅을 해왔으며; 결과적으로 우리는 훨씬 앞서 있다. 하지만 지금 중국을 무시하는 것은 2022년 말 ChatGPT가 나왔을 때 AI 경쟁에서 구글을 무시하는 것과 비슷하다. 구글은 아직 강도 높은 AI 베팅에 집중하지 않았고, OpenAI가 훨씬 앞서 있는 것처럼 보였지만—구글이 깨어난 1년 반 후, 그들은 매우 심각한 경쟁을 벌이고 있다. 중국 역시 매우 심각한 경쟁을 벌일 명확한 경로가 있다. 만약 중국공산당(CCP)이 AGI 경쟁에 동원된다면, 상황은 매우 달라질 수 있다.

어쩌면 중국 정부는 무능할 수도 있고; 어쩌면 AI가 CCP를 위협한다고 판단해 억압적인 규제를 부과할 수도 있다. 하지만 나는 그것에 기대지 않는다.

저는 개인적으로 우리가 전면적인 중국의 범용 인공지능(AGI) 노력을 마주하게 될 것이라는 가정 하에 움직여야 한다고 생각합니다. 매년 AI 능력이 극적으로 도약하고, 소프트웨어 엔지니어의 초기 자동화가 시작되며, AI 수익이 폭발적으로 증가하고 10조 달러 가치 평가와 조 단위 클러스터 구축이 나타나고, AGI가 임박했다는 광범위한 합의가 형성되기 시작할 때—중국공산당(CCP)은 이를 주목할 것입니다. 저는 이러한 도약들이 미국 정부(USG)를 AGI에 깨우칠 것으로 예상하는 만큼, CCP 역시 AGI에 눈을 뜨고 AGI에서 뒤처지는 것이 그들의 국가 권력에 어떤 의미인지 깨닫게 될 것으로 예상합니다.

그들은 강력한 적수가 될 것입니다.

권위주의적 위험

초지능의 힘을 휘두르는 독재자는 우리가 지금까지 본 적 없는 집중된 권력을 지니게 될 것입니다. 다른 나라에 자신의 의지를 강요할 수 있을 뿐만 아니라, 내부적으로도 자신의 통치를 확고히 할 수 있습니다. 수백만 대의 AI

107. 일부는 중국이 이 비밀들을 훔친다 해도 암묵적 지식이 필요하기 때문에 경쟁할 수 없다고 주장한다. 나는 동의하지 않는다. 나는 이것을 두 층으로 생각한다. 하단 층은 대규모 훈련 실행을 위한 엔지니어링 역량이다; 이 훈련 실행은 까다롭고 섬세하며 암묵적 지식을 필요로 한다. 하지만 나중에 논의하겠지만, 중국의 AI 노력은 대규모 모델을 훈련할 수 있음을 완벽히 보여주었고, 나는 그들이 이 암묵적 지식을 토착적으로 갖추고 있을 것이라 생각한다. 상단 층은 알고리즘 레시피—모델 아키텍처, 적절한 스케일링 법칙 등—으로, 이는 한 시간짜리 통화로 전달될 수 있다. 이러한 컴퓨팅 배수는 보통 불연속적인 변화이며, 따라서 대규모 훈련 실행을 위한 기본 암묵적 지식은 이전될 수 있다. 나는 “암묵적 지식”이 중국 AGI 노력에 결정적인 장벽이 되지 않을 것이라 본다.

통제 로봇 법 집행 요원이 국민을 감시할 수 있고, 대규모 감시는 극도로 강화되며, 독재자 충성 시가 각 시민을 개별적으로 평가하여 불충성을 뿌리 뽑는 고도화된 거의 완벽한 거짓말 탐지 기능을 갖출 수 있습니다.

가장 중요한 것은, 로봇 군대와 경찰력이 단일 정치 지도자에 의해 완전히 통제되고 완벽히 복종하도록 프로그래밍될 수 있다는 점입니다—더 이상 쿠데타나 대중 반란의 위험이 없습니다.

과거 독재 정권은 결코 영구적이지 않았지만, 초지능은 독재자의 통치에 대한 거의 모든 역사적 위협을 제거하고 그들의 권력을 고정시킬 수 있습니다(가치 고정 참조). 만약 CCP가 이 권력을 얻는다면, 그들은 당의 “진리” 개념을 완전하고 철저히 강제할 수 있을 것입니다.

분명히 말하자면, 저는 단지 “우리의 가치가 더 낮기 때문에” 독재자가 초지능을 갖는 것을 걱정하는 것이 아닙니다. 저는 자유와 민주주의를 강하게 믿습니다. 왜냐하면 제가 옳은 가치가 무엇인지 모르기 때문입니다. 역사의 긴 흐름 속에서, “시간은 많은 것을 뒤흔들어 놓았습니다.

“투쟁하는 신념들.” 나는 우리가 오류 수정, 실험, 경쟁, 적응의 메커니즘에 신뢰를 두어야 한다고 믿는다.

초지능은 그것을 다루는 이들에게 반대, 반대 의견을 짓밟고 인류를 위한 거대한 계획을 확고히 할 힘을 줄 것이다. 이 힘을 사용하는 끔찍한 유혹에 저항하기는 어려울 것이다. 나는 간절히 바라건대, 대신에 건국자들의 지혜에 의지하여 근본적으로 다른 가치들이 번성하게 하고, 미국 실험을 정의해온 시끄러운 다원성을 보존할 수 있기를 바란다.

범용 인공지능(AGI) 경쟁에서 걸려 있는 것은 단지 먼 대리 전쟁에서의 우위만이 아니라, 자유와 민주주의가 다음 세기 이상 살아남을 수 있느냐 하는 문제다. 인류 역사의 흐름은 잔인하면서도 명확하다. 20세기에 두 차례 전 세계를 위협한 폭정이 있었으며, 이 위협이 영원히 사라졌다고 착각해서는 안 된다. 많은 젊은 친구들에게 자유와 민주주의는 당연한 것으로 느껴지지만, 그렇지 않다. 역사상 가장 흔한 정치 체제는 권위주의다.¹⁰⁸

나는 중국공산당(CCP)과 그 권위주의 동맹들의 의도를 진심으로 알지 못한다. 하지만 상기할 점은: CCP는 아마도 인류 역사상 가장 위대한 전체주의 대량 학살자를 계속 숭배하는 정권이다(“기아, 박해, 강제 노동, 대량 처형으로 인한 희생자 수는 4천만에서 8천만 명으로 추정됨”); 최근 백만 명의 위구르인을 강제 수용소에 가두고 자유 홍콩을 짓밟은 정권; 신식(휴대폰 추적, DNA 데이터베이스, 얼굴 인식 등)과 구식(이웃을 신고하는 시민 군단 모집) 모두를 이용한 사회 통제를 위한 대규모 감시를 체계적으로 실행하는 정권; 모든 문자 메시지가 검열을 거치도록 하고, 해외에서 시위에 참여하는 자녀 때문에 가족을 경찰서로 끌고 가는 등 반대 의견을 억압하는 정권; 시진핑을 종신 독재자로 확립한 정권; 자유로운 이웃 국가를 군사적으로 짓밟고 “재교육”하겠다는 목표를 내세우는 정권; 중국 중심의 세계 질서를 명시적으로 추구하는 정권이다.

자유 진영은 이 경쟁에서 권위주의 세력을 반드시 이겨야 한다. 우리는 미국의 경제적·군사적 우위 덕분에 평화와 자유를 누리고 있다. 범용 인공지능(AGI)을 넘어 초지능으로 무장한 중국공산당(CCP)도 국제 무대에서 책임감 있게 행동하며 각자 영역을 존중할 수도 있다. 그러나 그들의 독재자 역사는 결코 아름답지 않다. 미국과 동맹국들이 이 경쟁에서 승리하지 못하면, 우리는 모든 것을 잃을 위험에 처하게 된다.

건전한 우위를 유지하는 것이 안전에 결정적이다

과학과 기술의 저주받은 역사는, 그들이 경이로움을 펼칠수록 파괴 수단도 함께 확장되었다는 것이다. 막대기와 돌에서부터 칼과 창, 소총과 대포, 기관총과 전차, 폭격기와 미사일, 핵무기에 이르기까지. 기술이 발전함에 따

라 ‘파괴력 대비 비용’ 곡선은 꾸준히 하락해왔다. 우리는 초지능 이후의 급속한 기술 진보도 이 추세를 따를 것으로 예상해야 한다.

아마도 생물학 분야의 극적인 발전은 조용하고 빠르게 퍼지며 명령에 따라 완벽한 치명성을 발휘하는 새로운 생물무기를 만들어낼 것이다(테러 단체도 감당할 수 있을 만큼 매우 저렴하게 제작 가능할 수도 있다). 아마도 새로운 유형의 핵무기는 핵무기 규모를 규모 차수(OOM) 단위로 증가시키고, 탐지 불가능한 새로운 운반 수단을 가능하게 할 것이다. 아마도 치명적인 독을 실은 모기 크기의 드론이 상대 국가의 모든 구성원을 표적으로 삼을 수도 있다. 한 세기 동안의 기술 진보가 무엇을 가져올지 알기 어렵지만, 끔찍한 가능성들이 펼쳐질 것임은 확실하다.

인류는 냉전 시절 자멸을 가까스로 피했다. 역사적 관점에서 볼 때, AGI가 제기하는 가장 큰 실존적 위험은 우리가 엄청난 대량 살상 수단을 개발할 수 있게 해준다는 점이다. 이번에는 이러한 수단이 악의적 행위자나 테러리스트에게까지 확산되어 접근 가능해질 수도 있다(특히 현재 경로대로라면, 초지능 모델 가중치가 충분히 보호받지 못해 북한, 이란 등에게 직접 도난당할 위험이 있다).

북한은 이미 조직적인 생물무기 프로그램을 보유하고 있다. 미국은 “북한이 생물무기를 개발·생산하기 위한 전담 국가 차원의 공격 프로그램을 가지고 있다”고 평가한다. 그들의 주요 제약은 소수의 최고 과학자들이 (합성) 생물학의 한계를 얼마나 밀어붙였는지에 달려 있는 것으로 보인다. 이 제약이 해소되어 수백만의 초지능을 활용해 생물무기 연구개발을 가속화할 수 있게 되면 어떻게 될까? 예를 들어, 미국은 북한이 현재 생물학적 제품을 유전공학적으로 조작할 수 있는 능력이 “제한적”이라고 평가한다. 이 능력이 무제한이 된다면 어떻게 될까? 어떤 불길한 신종 조합으로 우리를 인질로 잡게 될까?

더욱이, 초정렬(슈퍼얼라인먼트) 섹션에서 논의한 바와 같이, 지능 폭발 전후에는 극도의 안전 위험이 존재할 것이다. 우리는 초인적 AI 시스템을 신뢰하고 통제할 수 있도록 보장하는 새로운 기술적 도전에 직면하게 될 것이다. 이는 중요한 순간에 속도를 늦춰야 할 수도 있음을 의미한다. 예를 들어, 지능 폭발 중간에 6개월 정도 지연하여 안전성에 대한 추가 보장을 확보하거나, 컴퓨팅 자원의 상당 부분을 능력 향상보다는 정렬 연구에 투입하는 식이다.

일부는 안전에 관한 국제 조약을 기대하기도 한다. 그러나 나는 이것이 허황된 생각이라고 본다. 중국공산당(CCP)과 미국 정부(USG) 모두가 AGI에 대해 안전 위험을 진지하게 받아들일 만큼 준비된 세계는, 동시에 양측 모두 국제 경제 및 군사 우위가 걸려 있음을 인식하는 세계이기도 하다. AGI에서 몇 달 뒤쳐진다는 것은 영구적으로 뒤쳐지는 것을 의미할 수 있다. 만약 경쟁이 치열하다면, 적어도 초지능 초기 단계에서 어떤 군비 통제 균형도 매우 불안정할 것이다. 요컨대, “탈주(breakout)”는 너무 쉽다. 지능 폭발 경쟁에서 앞서 나가 초지능과 결정적 우위를 확보하려는 유인(그리고 다른 이들이 그렇게 할 것이라는 두려움)이 너무 크기 때문이다.¹⁰⁹ 최소한, 여기서 우리가 충분히 좋은 결과를 얻을 가능성은 희박해 보인다. (기후 조약들은 어떻게 되었나? 이 문제에 비하면 훨씬 쉬운 문제였던 것 같다.)

우리가 가진 주요—아마도 유일한—희망은 민주주의 국가들의 동맹이 적대 세력에 비해 건강한 우위를 확보하는 것이다.

108. 주로 군주제이다.

미국은 주도권을 잡아야 하며, 그 주도권을 이용해 전 세계에 안전 규범을 강제해야 한다. 이것이 우리가 핵무기 문제에서 취한 경로로, 평화적 핵기술 사용에 대한 지원을 제공하는 대신 국제 비확산 체제를 구축했고(궁극적으로는 미국 군사력에 의해 뒷받침됨), 이것이 유일하게 효과가 입증된 경로다.

아마도 가장 중요한 점은, 건강한 주도권이 우리에게 조정할 여지를 준다는 것이다. 예를 들어, 지능 폭발 시기에 정렬(alignment)에 추가적인 노력을 기울여 안전을 확보하기 위해 필요하다면 주도권의 일부를 ‘현금화’할 수 있는 능력이다.

초지능의 안전 문제는 만약 치열한 군비 경쟁 상태라면 관리하기 매우 어려워진다. 2년과 2개월의 주도권 차이는 쉽게 모든 차이를 만들 수 있다. 만약 우리가 2개월의 주도권만 가진다면, 안전을 위한 여유가 전혀 없다. 중국공산당(CCP)의 지능 폭발을 두려워하여 우리는 거의 틀림없이 무제한으로 우리 자신의 지능 폭발을 경쟁할 것이다. 몇 달 만에 인간보다 훨씬 똑똑한 AI 시스템을 향해 돌진하면서, 핵심 결정을 올바르게 내릴 여유 없이, 초지능이 잘못될 위험을 모두 감수하게 된다. 우리는 매우 불안정한 상황에 직면할 것이며, 우리와 CCP가 빠르게 새로운 군사 기술을 개발하면서 역지력을 반복적으로 불안정하게 만들 것이다. 만약 우리의 비밀과 모델 가치가 잠겨 있지 않다면, 다른 여러 불량 국가들도 초지능을 이용해 자신들의 새로운 초대형 WMD 무기고를 갖추는 데 근접할 수 있다. 우리가 간신히 앞서 나간다고 해도, 그것은 아마도 피로 얻은 승리일 것이다; 존재론적 투쟁이 세계를 완전한 자기 파괴 직전까지 몰고 갈 것이다.

민주주의 동맹국들이 건강한 주도권, 예를 들어 2년의 주도권을 가진다면 초지능은 매우 다르게 보인다.¹¹⁰ 이것은 우리가 초지능을 둘러싼 전례 없는 일련의 도전을 헤쳐 나가고 상황을 안정시키는 데 필요한 시간을 확보해 준다.

미국이 결정적으로 승리할 것이 분명해지면, 그때 중국과 다른 적대국들에게 협상을 제안한다. 그들은 자신들이 이길 수 없다는 것을 알게 되고, 따라서 그들의 유일한...

선택지는 협상 테이블에 나오는 것이다; 그리고 우리는 그들이 서구의 노력을 방해하기 위한 열광적인 대치나 막판 군사적 시도를 피하기를 바란다. 그들의 내정 불간섭을 보장하고, 초지능의 평화적 혜택을 공유하는 대가로, 비확산 체제, 안전 규범, 그리고 초지능 이후의 안정성의 외형이 탄생할 수 있다.

어쨌든, 이 투쟁이 깊어질수록 우리는 자기 파괴의 위협을 잊어서는 안 된다. 우리가 냉전 시기를 무사히 넘긴 것은 너무나도 큰 행운¹¹¹이었다 — 그리고 그 파괴력은 당시 우리가 직면했던 것보다 천 배나 강력할 수 있다. 미국 주도의 민주주의 연합이 건강한 우위를 점하고 — 우리가 처한 어떤 불안정한 상황이든 이를 안정시키기 위한 엄숙한 리더십 행사가 — 아마도 이 벼랑을 넘어가는 가장 안전한 길일 것이다. 하지만 AGI 경쟁의 열기 속에서, 우리는 그것을 망쳐서는 안 된다.

109. 다음 비교를 고려해 보라. 불안정한 군비 통제 균형과 안정적인 군비 통제 균형의 차이. 1980년대 냉전 시기 군비 통제는 핵무기를 상당히 줄였지만 안정적인 균형을 목표로 했다. 미국과 소련은 여전히 수천 기의 핵무기를 보유했다. 어느 한쪽이 핵무기를 급히 증강하려 해도 상호확증파괴(MAD)가 보장되었다. 불량 국가가 핵무기를 일부 개발하려 해도 미국이나 소련을 근본적으로 위협할 수는 없었다. 그러나 무기 수를 매우 낮은 수준으로 감축하거나 급격한 기술 변화가 일어나는 상황에서는 균형이 불안정하다. 불량 행위자나 조약 위반자가 급히 핵무기 증강 프로그램을 시작해 다른 국가들을 완전히 압도할 수 있다. 핵무기 제로는 안정적인 균형이 될 수 없다. 이 논문은 1차 세계대전 이후 군비 제한과 워싱턴 해군 조약 같은 역사적 사례를 통해, 유사하게 역동적인 상황에서 군비 감축이 안정화가 아니라 불안정을 초래했음을 보여준다. 만약 AGI에서 몇 달의 우위가 결정적인 이점을 준다면, AI 분야에서의 안정적인 군비 감축도 마찬가지로 어렵다. 불량 신흥 세력이나 조약 위반자가 비밀리에 급히 프로그램을 시작해 큰 우위를 차지할 수 있으며, 어떤 합의도 안정적일 수 없을 것이다.

110. 이미 오늘날 AI 발전 속도가 매우 빠르고, 지능 폭발과 초지능 이후의 더 광범위한 기술 폭발에서 기대되는 더욱 빠른 속도를 고려할 때, “단지” 2년의 주도권도 초지능 이후 세계에서 능력 차이에 있어 엄청난 차이가 될 것이다.

초지능은 국가 안보 문제다

명백하다: AGI는 미국 국가 안보에 대한 존재론적 도전이다. 이제 이를 그렇게 다루기 시작할 때다.

미국 정부는 천천히 움직이기 시작하고 있다. 미국 칩에 대한 수출 통제는 매우 중요한 사안이며, 당시에는 놀라울 정도로 선견지명이 있는 조치였다. 하지만 우리는 전 분야에서 진지해져야 한다.

미국은 우위를 점하고 있다. 우리는 그 우위를 유지하기만 하면 된다. 그런데 지금 우리는 그것을 망치고 있다. 무엇보다도, 우리는 향후 12~24개월 내에 주요 AGI 돌파구(또는 AGI 가중치 자체)가 유출되기 전에 AI 연구소를 신속하고 근본적으로 봉쇄해야 한다. 우리는 독재국가들이 쉽게 돈을 제공하는 곳이 아니라 미국 내에 컴퓨터 클러스터를 구축해야 한다. 그리고 그렇다, 미국 AI 연구소는 정보기관 및 군과 협력할 의무가 있다. 미국의 AGI 우위는 단지 최고의 AI 여자친구 앱을 만드는 것으로 평화와 자유를 확보하지 못할 것이다. 보기 좋지는 않지만 — 우리는 미국 방어를 위해 AI를 구축해야 한다.

우리는 이미 수십 년 만에 가장 폭발적인 국제 정세를 향해 나아가고 있다. 푸틴은 동유럽에서 진격 중이다. 중동은 불타고 있다. 중국공산당(CCP)은 대만을 자신의 운명으로 보고 있다. 여기에 범용 인공지능(AGI) 경쟁을 더하라. 초지능 이후 수년 내에 한 세기 분량의 기술 혁신이 압축되어 일어날 것이다. 이는 지금까지 본 것 중 가장 불안정한 국제 정세 중 하나가 될 것이며, 적어도 초기에는 선제공격에 대한 유인이 엄청날 것이다.¹¹² 모두가 자신만의 초지능을 향해 달려갈 것이며, 누군가가 돌이킬 수 없이 앞서 나가기 전까지는 제한된 시간이 있을 것이다. 적의 초지능 클러스터를 충분한 물리적 우위를 확보하기 전에 무력화하려는 큰 유인이 있을 것이다(예: 초지능을 이용해 뚫을 수 없는 미사일 방어 체계나 드론 무리를 개발하는 것). 이로 인해 다른 모든 이들은 영원히 뒤처지게 된다.

AGI 예상 시기(~2027년?)와 대만 침공 예상 시기(중국이 2027년까지 대만 침공 준비 완료?)가 이미 오싹할 정도로 일치하고 있다—이 일치하는 세계가 AGI를 인식하면서 더욱 심화될 것이다. (1960년에 세계 대부분의 우라늄 매장지가 베를린에 집중되어 있었다고 상상해 보라!) 내 생각에 AGI 최종 국면이 세계 대전이라는 배경 속에서 전개될 현실적인 가능성이 있다. 그때는 모든 내기가 무의미해진다.

IV. 프로젝트

범용 인공지능(AGI) 경쟁이 치열해짐에 따라 국가안보 체제가 개입할 것이다. 미국 정부(USG)는 잠에서 깨어나 27년 또는 28년쯤에는 어떤 형태로든 정부 주도의 AGI 프로젝트가 시작될 것이다. 어떤 스타트업도 초지능을 다룰 수 없다. 민감구획정보시설(SCIF) 어딘가에서 최종 국면이 펼쳐질 것이다.

우리는 수백 개의 발전소, 수천 개의 폭탄, 수만 명의 사람들이 국가 기관에 집결한 이와 같은 집합체가 어떻게 실험실 벤치에 앉아 한 종류의 원자 특이한 행동을 논의하는 몇몇 사람들로부터 유래했는지 알고자 하는 호기심을 가져야 한다. 스펜서 R. 위어트

111. 다니엘 엘스버그는 당시 RAND와 국가안보 기구에서 핵전쟁 계획자로 일하며 이 이야기를 흥미진진하게 전한다.

112. 모두가 자신만의 초지능을 향해 경쟁하며, 한쪽이 되돌릴 수 없이 앞서나가기 전 제한된 기간만 존재합니다. 일단 충분히 물리적 우위를 확보하기 전 경쟁 상대의 초지능 클러스터를 무력화하려 할 큰 유인도 있을 것입니다(예: 초지능으로 요격 불가능 미사일 방어체계나 드론 떼 개발).

요즘에는 최첨단 AI 시스템의 라이선스 발급부터 안전 기준, 학계에 수억 단위의 컴퓨팅을 제공하는 공용 클라우드에 이르기까지 다양한 “AI 거버넌스” 계획이 제안되고 있다. 이들은 선의로 보이지만, 내게는 범주 오류를 범하는 것처럼 보인다.

미국 정부가 무작위 샌프란시스코 스타트업에게 초지능 개발을 맡길 것이라는 생각은 미친 짓이다. 만약 우버에게 즉흥적으로 원자폭탄을 개발하게 했다면 어땠을지 상상해 보라.

초지능—인간보다 훨씬 똑똑한 AI 시스템—은 신무기 개발에서부터 경제 성장 폭발을 이끄는 데 이르기까지 막대한 힘을 가질 것이다. 초지능은...

국제 경쟁의 중심지; 군사 충돌에서 결정적일 수 있는 몇 달의 선도.

역사로부터의 우리의 짧은 휴식을 무의식적으로 내면화한 이들의 착각은 이것이 더 원초적인 힘들을 불러일으키지 않을 것이라는 점이다. 우리 이전의 많은 과학자들처럼, 샌프란시스코의 위대한 지성들은 자신들이 탄생시키고 있는 악마의 운명을 통제할 수 있기를 바란다. 지금은 그들이 여전히 통제할 수 있다; 왜냐하면 그들은 상황 인식을 가진 소수 중 하나로, 자신들이 무엇을 만들고 있는지 이해하고 있기 때문이다. 그러나 앞으로 몇 년 안에, 세계는 깨어날 것이다. 국가안보 체제도 마찬가지다. 역사는 승리의 귀환을 할 것이다.

과거 여러 차례처럼—코로나, 제2차 세계대전—미국이 운전대를 놓친 듯 보이다가도, 갑자기 정부가 가장 비범한 방식으로 움직이기 시작할 것이다. 몇 년 안에, 단지 몇 차례 더 “2023년 수준”의 모델 능력과 AI 담론의 도약이 있을 때, 분명해질 순간이 올 것이다: 우리는 범용 인공지능(AGI)의 문턱에 있으며, 곧 초지능이 뒤따를 것이다. 정확한 메커니즘에는 많은 변동이 있지만, 어쨌든 미국 정부가 지휘봉을 잡을 것이다; 주요 연구소들은 (“자발적으로”) 합병할 것이며; 의회는 칩과 전력에 수조 달러를 배정할 것이다; 민주주의 국가들의 연합이 형성될 것이다.

스타트업은 많은 일에 훌륭하지만—스타트업 단독으로는 미국의 가장 중요한 국가 방위 프로젝트를 책임질 준비가 되어 있지 않다. 우리는 직면할 전면적인 첩보 위협에 맞서기 위해 정부의 개입이 필요할 것이다; 민간 AI 노력은 사실상 초지능을 중국공산당(CCP)에 직접 전달하는 것과 다름없다. 우리는 최소한의 정상적인 지휘 체계라도 확보하기 위해 정부가 필요하다; 무작위 CEO(또는 무작위 비영리 이사회)가 핵 버튼을 가질 수는 없다. 우리는 초지능의 심각한 안전 문제를 관리하고, 지능 폭발의 전쟁의 안개를 관리하기 위해 정부가 필요하다. 우리는 초지능을 배치하여 어떤 극단적 위협이 발생하더라도 방어하고, 이후 이어질 극도로 불안정하고 불안정한 국제 상황을 극복하기 위해 정부가 필요하다. 우리는 민주주의 연합을 동원하여 경쟁에서 승리하기 위해 정부가 필요할 것이다.

권위주의 세력과 협력하여 전 세계를 위한 비확산 체제를 구축(및 집행)해야 한다. 이렇게 되어야 한다는 점이 안타깝지만—정부가 필요할 것이다. (네, 행정부가 누구든 상관없이.)

어쨌든, 나의 주된 주장은 규범적이기보다는 기술적이다. 몇 년 안에 프로젝트가 시작될 것이다.

프로젝트로 가는 길

내 기억에 선명히 각인된 사건은 2020년 2월 말부터 3월 중순까지다. 2월 마지막 주와 3월 초, 나는 완전히 절망에 빠져 있었다. 우리는 코로나의 지수적 확산 단계에 접어든 것이 분명해 보였고, 전염병이 나라를 휩쓸 것이며 병원 붕괴가 임박했는데도 거의 아무도 심각하게 받아들이지 않았다. 뉴욕 시장은 여전히 코로나 공포를 인

종차별로 치부하며 사람들이 브로드웨이 공연에 가도록 권장했다. 내가 할 수 있는 것은 마스크를 사는 것과 주식을 공매도하는 것뿐이었다.

그런데 불과 몇 주 만에 전국이 봉쇄되었고 의회는 수조 달러(실제로 GDP의 10% 이상)를 배정했다. 지수적 확산이 어디까지 갈지 미리 예측하는 것은 너무 어려웠지만, 위협이 충분히 가까워지고 존재론적으로 심각해지자 비상한 힘들이 발휘되었다. 대응은 늦고 조잡하며 무딘 것이었지만, 어쨌든 이루어졌고 극적이었다.

앞으로 몇 년간 AI 분야도 비슷한 느낌일 것이다. 우리는 지금 중반전이다. 2023년은 이미 격변의 해였다. AGI는 주저하며 연관 짓던 주변 주제에서 주요 상원 청문회와 세계 정상 회담의 주제로 급부상했다. 아직 초기 단계임에도 불구하고 미국 정부의 참여 수준은 나에게 인상적이었다. 2023년과 같은 해가 몇 번 더 지나면 오버턴 창(Overton window)이 완전히 열릴 것이다.

우리가 규모 차수(OOM)를 빠르게 넘나들면서 도약은 계속될 것이다. 2025년이나 2026년쯤에는 다음으로 충격적인 단계적 변화가 일어날 것으로 예상된다; AI는 대형 기술 기업에 연간 1000억 달러 이상의 수익을 창출하고, 박사 학위 소지자들을 순수 문제 해결 능력에서 능가할 것이다. 마치

코로나로 인한 주식시장 붕괴는 많은 이들로 하여금 코로나를 심각하게 받아들이게 만들었고, 우리는 10조 달러 규모의 기업들을 보게 될 것이며 AI 열풍이 어디에나 퍼질 것이다. 그것만으로도 부족하다면, 2027년/28년쯤에는 1000억 달러 이상 규모의 클러스터에서 훈련된 모델들이 등장할 것이다; 본격적인 AI 에이전트/원격 근무자가 소프트웨어 엔지니어링과 기타 인지적 업무를 광범위하게 자동화하기 시작할 것이다. 매년 그 가속도는 현기증이 날 정도로 느껴질 것이다.

많은 이들이 아직 범용 인공지능(AGI)의 가능성을 보지 못하지만, 결국 합의가 형성될 것이다. 어떤 이들은, 예를 들어 실라드(Szilard)처럼, 다른 이들보다 훨씬 일찍 원자폭탄의 가능성을 보았다. 그들의 경고는 처음에는 잘 받아들여지지 않았고; 폭탄의 가능성은 먼 이야기로 치부되었으며(적어도 보수적이고 올바른 태도는 그 가능성을 축소하는 것이었다고 여겨졌다). 실라드의 열렬한 비밀 유지 호소는 조롱당하고 무시되었다. 그러나 점점 더 많은 실험적 결과가 나오면서, 처음에는 회의적이던 많은 과학자들이 폭탄이 가능하다는 것을 깨닫기 시작했다. 과학자 대다수가 우리가 폭탄의 문턱에 서 있다고 믿게 되자, 정부는 국가안보의 긴급성을 너무 크게 인식했고—맨해튼 프로젝트가 시작되었다.

규모 차수(OOM)가 이론적 추정에서 (비범한) 실증적 현실로 전환됨에 따라, 점차 주요 과학자들과 경영진, 정부 관리들 사이에서도 합의가 형성될 것이다: 우리는 범용 인공지능(AGI), 지능 폭발, 초지능의 문턱에 서 있다. 그리고 이 즈음에 우리는 AI의 진정으로 무서운 첫 시연을 보게 될 것이다: 아마도 자주 논의되는 “초보자들이 생물무기를 만드는 것을 돕는” 사례, 혹은 자율적으로 중요한 시스템을 해킹하는 사례, 또는 전혀 다른 무언가일 것이다. 이 기술이 완전히 결정적인 군사 기술임이 분명해질 것이다. 우리가 큰 전쟁에 휘말리지 않는 행운을 누리더라도, 중국공산당(CCP)이 이를 주목하고 강력한 AGI 노력을 시작했을 가능성이 높다. 아마도 결국 (불가피한) CCP가 미국의 주요 AI 연구소에 침투한 사실이 밝혀져 큰 파장을 일으킬 것이다.

대략 26년/27년 무렵, 워싱턴의 분위기는 엄숙해질 것이다. 사람들은 무슨 일이 일어나고 있는지 본능적으로 느끼기 시작할 것이며; 두려워할 것이다. 펜타곤의 복도에서부터 비공개 의회 브리핑실까지 명백한 소리가 울려 퍼질 것이다.

모두가 궁금해하는 질문: 우리는 범용 인공지능(AGI) 맨해튼 프로젝트가 필요한가? 처음에는 천천히, 그러다 갑자기 모두가 알게 될 것이다: 이 일이 일어나고 있으며, 상황이 급격히 변할 것이고, 이것이 원자폭탄 발명 이후

미국 국가안보에 가장 중요한 도전이라는 것을. 어떤 형태로든 국가안보 체제가 매우 깊이 관여하게 될 것이다. 이 프로젝트는 필수적이며, 사실상 유일하게 그럴듯한 대응책이 될 것이다.

물론, 이것은 매우 간략한 설명이다—많은 것은 언제, 어떻게 합의가 형성되는지, 주요 경고 신호 등이 좌우한다. 워싱턴 DC는 악명 높게 비효율적이다. 코로나 사태나 심지어 맨해튼 프로젝트 때처럼, 정부는 믿을 수 없을 정도로 늦고 서투를 것이다. 1939년 아인슈타인이 대통령에게 보낸 편지(질라드가 작성) 이후 우라늄 자문위원회가 구성되었지만, 관료들은 무능했고 초기에는 별다른 진전이 없었다. 예를 들어, 페르미는 연구 지원금으로 겨우 6천 달러(현재 가치 약 13만 5천 달러)를 받았는데, 그것도 쉽게 주어진 것이 아니었고 몇 달을 기다려야 했다. 질라드는 당국의 근시안적이고 느린 대응 때문에 프로젝트가 최소 1년 지연되었다고 믿었다. 1941년 3월, 영국 정부는 결국 폭탄 개발이 불가피하다고 결론지었다. 미국 위원회는 이 영국 보고서를 몇 달간 완전히 무시하다가, 결국 1941년 12월에야 전면적인 원자폭탄 개발 노력을 시작했다.

이것이 실제로 어떻게 운영될지는 여러 가지 방법이 있을 수 있다. 분명히 말하자면, 이것이 반드시 AI 연구자들이 군에 고용되는 식의 문자 그대로의 국유화일 필요는 없다(물론 그럴 수도 있다!).¹¹³ 오히려 나는 더 세련된 조율 방식을 예상한다. 국방부(DoD)와의 관계는 보잉이나 록히드 마틴과의 관계와 비슷할 수 있다. 주요 클라우드 컴퓨팅 제공업체, AI 연구소, 정부 간에 방위 계약이나 유사한 형태로 합작 투자가 이루어져, 사실상 국가안보 체제의 프로젝트가 되는 것이다. 2023년에 AI 연구소들이 백악관에 “자발적으로” 약속을 한 것처럼, 서방 연구소들도 어느 정도 “자발적으로” 국가적 노력에 합류할 수 있다. 그리고 수조 달러의 투자가 관련되므로, 견제와 균형을 위해 의회가 관여해야 할 가능성이 크다.¹¹⁴

이 모든 세부 사항이 어떻게 정리될지는 다음 기회에 다룰 이야기다.

하지만 26/27/28년 말까지는 진행 중일 것이다. 핵심 범용 인공지능(AGI) 연구팀(수백 명의 연구원)은 안전한 장소로 이동할 것이며, 조 단위 달러 규모의 클러스터가 기록적인 속도로 구축될 것이다; 프로젝트는 시작될 것이다.

왜 프로젝트가 유일한 방법인가

나는 정부에 대해 환상을 가지고 있지 않다. 정부는 온갖 제약과 잘못된 인센티브에 직면해 있다. 나는 미국 민간 부문을 크게 신뢰하며, 기술이나 산업에 정부가 과도하게 개입하는 것을 거의 권장하지 않는다.

나는 AGI에도 같은 틀을 적용했었다—AI 연구소에 합류하기 전까지는. AI 연구소는 몇 가지 면에서 매우 뛰어나다: AI를 학문적 과학 프로젝트에서 스타트업만이 할 수 있는 방식으로 상업적 대무대로 끌어올렸다. 하지만 궁극적으로 AI 연구소는 여전히 스타트업이다. 우리는 스타트업이 초지능을 다룰 준비가 되어 있다고 기대해서는 안 된다.

여기에는 좋은 선택지가 없다—하지만 다른 방법도 보이지 않는다. 기술이 국가 안보에 이토록 중요해지면, 우리는 미국 정부(USG)가 필요할 것이다.

113. 참고로, 민간 기업들이 핵무기 부품 개발에 참여하지만 완성된 핵무기를 소유하는 것은 절대 허용되지 않는다. 이에 비해 내가 제안하는 “AGI 정부 프로젝트”의 주류 버전은 대량살상무기(WMD) 참조 범주에서 전례 없이 민간화되어 있다.

114. 의회—심지어 부통령조차도—맨해튼 프로젝트에 대해 알지 못했다. 우리는 여기서 그런 실수를 반복하지 말아야 하며, 프로젝트의 핵심 관계자들은 상원의 인준을 받아야 한다고 제안한다.

초지능은 미국의 가장 중요한 국가 방위 프로젝트가 될 것이다

나는 이전 글에서 초지능의 힘에 대해 논의했다. 수년 내에 초지능은 군사력 균형을 완전히 뒤흔들 것이다. 2030년대 초까지는 미국의 모든 무기(좋은 싫든, 세계 평화와 안보의 기반)가 아마도 구식이 될 것이다. 이는 단순한 현대화 문제가 아니라 전면적인 교체가 될 것이다.

간단히 말해, AGI 개발은 인터넷보다는 핵무기와 더 비슷한 범주에 속하게 될 것이 분명하다. 물론 이중 용도 기술이겠지만—핵 기술도 이중 용도였다. 민간 응용 분야도 그 시기를 가질 것이다. 하지만...

AGI 종말의 안개 속에서, 좋은 나쁜 국가안보가 주요 배경이 될 것이다.

우리는 급격한 기술 변화에 직면하여 수년 내에 미국 군대를 완전히 재편해야 할 것이다—그렇지 않으면 그렇게 하는 적들에게 완전히 압도당할 위험이 있다. 아마도 가장 중요한 초기 우선순위는 방어적 응용을 위해 초지능을 배치하는 것이며, 무수히 많은 새로운 위협을 견뎌내기 위한 대응책을 개발하는 것이다: 초인적인 해킹 능력을 가진 적들, 우리의 핵 역지력을 선제 공격할 수 있는 새로운 유형의 은밀한 드론 무리, 무기화될 수 있는 합성 생물학의 진보 확산, 격동하는 국제(및 국내) 권력 투쟁, 그리고 불량 초지능 프로젝트들.

명목상 민간이든 아니든, AGI 프로젝트는 반드시, 그리고 실제로는 방위 프로젝트가 되어야 하며, 국가안보 체제와 매우 긴밀한 협력이 필요할 것이다.

초지능을 위한 건전한 지휘 체계

초지능의 힘과 도전은 우리가 기술 기업에서 보아온 어떤 것과도 매우 다른 참조 범주에 속할 것이다. 분명해 보인다: 이것은 무작위 CEO의 일방적 지휘 하에 있어서는 안 된다. 실제로, 초지능을 개발하는 민간 연구소 세계에서 개별 CEO들이 문자 그대로 미국 정부를 쿠데타할 권력을 가질 가능성도 충분하다.¹¹⁵ 엘론 머스크가 핵 무기 지휘권을 최종적으로 갖는 상황을 상상해 보라.¹¹⁶ (또는 무작위 비영리 이사회가 핵 무기 지휘권을 장악하기로 결정하는 경우도 마찬가지다.)

사회로서 우리는 군대를 민주 정부가 통제해야 한다고 결정했다는 점은 아마도 명백하다;¹¹⁷ 초지능은 적어도 처음에는 가장 강력한 군사 무기가 될 것이다. 급진적인 제안은 프로젝트 자체가 아니라, 민간 AI CEO들이 군사 권력을 휘두르며 자비로운 독재자가 될 것이라는 내기에 베팅하는 것이다.

(실제로 민간 AI 연구소 세계에서 핵 버튼을 가진 무작위 CEO보다 상황이 더 나쁠 가능성이 크다—AI의 일부가)

연구소들의 끔찍한 보안 문제는 내부 통제의 전무함에 있다. 즉, 무작위로 선발된 AI 연구소 직원들(전혀 검증되지 않은)이 눈치채이지 않은 채 무단 행동을 할 수 있다는 것이다.

우리는 합리적인 지휘 체계가 필요하다—그리고 WMD(대량살상무기)에 비견될 만한 것을 책임감 있게 다루기 위해 반드시 수반되는 모든 절차와 안전장치도 필요하다—그리고 이는 정부가 수행해야 할 일이다. 어떤 면에서

115. 이 시점에서는 AI 연구소 직원의 협력이 필요하지 않을 것입니다. 대부분 자동화되어 있을 것이기 때문입니다.

116. Sam Altman이 “매년 AGI에 가까워질수록 모두 +10의 미친 점수를 얻는다”고 말한 바 있습니다.

117. 실제로 정부가 가장 무거운 무기를 지닌 것은 거대한 문명적 성취였습니다! 중세 같은 ‘모두가 모두와 싸우는’ 상태 대신, 법원과 다원적 기관을 통해 분쟁을 해결할 수 있었습니다.

이것은 단순히 버크주의적 주장이다: 정부 권력을 견제하는 제도, 헌법, 법률, 법원, 견제와 균형, 규범, 자유민주주의 질서에 대한 공통된 헌신(예: 장군들이 불법 명령을 거부하는 것) 등이 수백 년의 시험을 견뎌냈다. 반면, 특별한 AI 연구소 거버넌스 구조는 처음 시험대에 올랐을 때 무너졌다. 미국 군대는 이미 원한다면 미국 내 거의 모든 민간인을 죽이거나 권력을 장악할 수 있다—그리고 우리가 핵무기에 대한 정부 권력을 견제하는 방식은 각기 핵무기를 보유한 수많은 민간 기업을 통해서가 아니다. 이 임무를 수행할 수 있음이 입증된 지휘 체계와 제도는 단 하나뿐이다.

다시 말하지만, 당신이 진정한 자유지상주의자라서 규범적으로 동의하지 않을 수도 있다(일론 머스크와 샘 알트먼이 자신들의 핵무기를 지휘하게 하라!).¹¹⁸ 그러나 초지능이 국가 안보의 핵심 문제임이 명확해지면, 워싱턴 DC의 남녀들은 분명 그렇게 볼 것이다.

초지능의 민간 활용

물론, 초지능의 민간 활용이 정부에만 국한된다는 의미는 아니다.

- 핵 연쇄 반응은 처음에 정부 프로젝트로 활용되었고, 핵무기는 영구히 정부에만 허용되었지만, 민간 핵 에너지는 60~70년대에 민간 프로젝트로 번성했다(환경운동가들이 중단시키기 전까지).
- 보잉은 B-29를 만들었다(제2차 세계대전 중 가장 비싼 방위 연구개발 프로젝트로, 맨해튼 프로젝트보다 더 비쌌다).

어쨌든, AGI에 가까워질수록 오픈소스의 중요성을 과대평가하는 경향이 있다. 클러스터 비용이 수천억 달러로 치솟고, 주요 알고리즘 비밀이 몇 년 전과 달리 공개되지 않고 독점적으로 관리되기 때문에, AGI를 개발하는 주체는 몇몇 선도 기업 2~3곳이 될 것이지, 분산된 코더들의 행복한 커뮤니티가 아닐 것이다.

나는 다른 형태의 오픈소스가 계속 중요한 역할을 할 것이라고 생각한다: 몇 년 뒤처진 모델들이 오픈소스로 공개되어 기술의 혜택이 널리 확산되는 데 기여하는 것이다.

(프로젝트)와 B-47 및 B-52 장거리 폭격기를 군과 협력하여 개발한 후, 그 기술을 보잉 707 상업용 비행기에 적용하여 제트 시대를 열었다. 그리고 오늘날 보잉은 스텔스 전투기를 정부에만 판매할 수 있지만, 민간 제트기는 자유롭게 개발하고 판매할 수 있다.

- 레이더, 위성, 로켓, 유전자 기술, 제2차 세계대전 공장 등도 마찬가지였다.

초지능의 초기 개발은 극도로 불안정한 시기를 생존하고 안정화하려는 국가안보의 긴급성에 의해 주도될 것이다. 초지능의 군사적 사용은 정부에 한정될 것이며, 안전 규범이 시행될 것이다. 그러나 초기 위기가 지나면, 국가 컨소시엄에 참여한 기업들(및 기타 기업들)이 민간 응용 분야를 사적으로 추구하는 것이 자연스러운 경로가 될 것이다.

프로젝트가 존재하는 세계에서도, 초지능의 민간 응용 분야는 사적이고 다원적이며 시장 기반의 번성하는 생태계를 갖게 될 것이다.

118. 또는 모든 것을 오픈소스로 공개하자고 말할 수도 있다. 단순히 모든 것을 오픈소스로 공개하는 문제는 미국 내에서 수많은 꽃이 피는 행복한 세상이 아니라, 중국공산당(CCP)이 미국이 개발한 초지능에 자유롭게 접근하여 더 많이 만들고(그리고 덜 조심하거나 규제하며) 세계를 장악하는 세상이 된다는 점이다. 그리고 또 다른 문제는 초대형 대량살상무기가 전 세계의 모든 불량 국가와 테러 단체에 확산된다는 점이다. 나는 이것이 좋은 결말을 맺지 못할 것이라고 본다. 정부가 전혀 없는 상태가 자유보다 오히려 폭정(또는 파괴)로 이어질 가능성이 더 높은 것 과 비슷하다.

보안

이 문제에 대해서는 이전 시리즈에서 길게 다룬 바 있다. 현재의 경로라면, 미국의 범용 인공지능(AGI) 노력을 포기하는 것이나 다름없다; 중국공산당(CCP)은 모든 알고리즘 혁신과 모델 가중치(말 그대로 초지능의 복사본)를 즉시 훔칠 수 있다. 현재 경로에서는 초지능에 대해 ‘북한 방어’ 수준의 보안을 확보할 수 있을지도 불분명하다. 민간 스타트업이 AGI를 개발하는 세계에서는 초지능이 수십 개의 불량 국가로 확산될 것이다. 이는 단순히 용납할 수 없다.

이 문제를 진지하게 다룰 생각이라면, 분명히 이 기술을 철저히 통제해야 한다. 대부분의 민간 기업들은 이를 심각하게 받아들이지 못했다. 어쨌든, 결국 중국의 첩보 활동(예: 가중치 탈취가 국가안전부(MSS)의 최우선 과제임)을 온전히 막아내려면, 민간 기업이 충분한 보안을 확보하는 것은 아마 불가능할 것이다. 이는 요구될 것이다.

그 시점에서 AGI를 충분히 안전하게 확보하기 위해서는 미국 정보기관의 광범위한 협력이 필요할 것이다. 이는 AI 연구소와 AGI 핵심 연구진에 대한 극단적인 신원 조사, 지속적인 감시, 민감구획정보시설(SCIF) 내 근무, 출입 자유 제한 등 침해적인 제약을 포함하며, 궁극적으로는 정부만이 제공할 수 있는 인프라, 즉 AGI 데이터센터 자체의 물리적 보안도 요구할 것이다.

어떤 면에서는 보안만으로도 정부 프로젝트가 필수적임을 입증한다—자유 진영의 우위와 AI 안전 모두 이 기술을 잠글 수 없다면 실패할 것이다. (사실, 중국공산당(CCP)의 AGI 연구소 침투가 명확해지면 모든 상원의원, 하원의원, 국가안보 관계자들이 이 문제에 대해 강한 의견을 갖게 될 가능성이 상당히 높다고 생각한다.)

안전

간단히 말해: 우리가 이 문제를 망칠 수 있는 방법은 많다—곧 우리 경제와 군대를 책임질 수십억의 초지능 에이전트를 신뢰하고 통제할 수 있도록 보장하는 것(조정될 문제)부터, 새로운 대량 파괴 수단의 오용 위험을 통제하는 것까지.

일부 AI 연구소는 안전에 전념한다고 주장한다: 그들이 만드는 것이 잘못되면 재앙을 초래할 수 있음을 인정하고, 때가 되면 필요한 조치를 취하겠다고 약속한다. 나는 그들의 약속을 믿고 모든 미국인의 생명을 걸 수 있을지 모르겠다. 더 중요한 것은, 지금까지 그들이 스스로 인정하는 바를 구축하는 데 필요한 역량, 신뢰성, 진지함을 입증하지 못했다는 점이다.

본질적으로 그들은 스타트업이며, 모든 일반적인 상업적 인센티브를 갖고 있다. 경쟁은 모두가 지능 폭발을 단순히 앞다투어 달성하도록 몰아갈 수 있으며, 적어도 일부 행위자는 안전을 무시할 의지가 있을 것이다. 특히, 우리는 안전 문제를 해결할 시간을 벌기 위해 “우리가 가진 우위를 일부 사용”하고자 할 수 있지만, 이를 위해서는 서방 연구소들이 협력해야 한다. (물론, 민간 연구소들은 이미 그들의 AGI 모델 가중치가 도난당했을 것이므로, 그들의 안전 준비는...)

주의 사항은 전혀 중요하지 않을 수도 있다; 우리는 중국공산당(CCP)과 북한의 안전 조치에 좌우될 것이다.)

한 가지 해답은 규제이다. AI가 더 천천히 발전하는 세상에서는 적절할 수 있지만, 나는 규제가 지능 폭발이라는 도전의 본질에 맞지 않을 것이라고 두려워한다. 필요한 것은 몇 년 동안 신중한 평가를 하고 관료주의를 통해 몇 가지 안전 기준을 통과시키는 것과는 다를 것이다. 그것은 전쟁을 치르는 것과 더 비슷할 것이다.

우리는 매주 상황이 극도로 빠르게 변하는 미친 한 해를 맞이할 것이다. 모호한 데이터에 근거한 어려운 결정이 생사와 직결될 것이며, 해결책—심지어 문제 자체도—사전에 완전히 명확하지 않고 ‘전쟁의 안개’ 속에서의 역량에 달려 있을 것이다. 여기에는 “우리의 일부 정렬 측정치가 모호하게 보이고, 무슨 일이 일어나고 있는지 더 이상 잘 이해하지 못한다. 괜찮을 수도 있지만, 다음 세대 초지능이 잘못될 수 있다는 경고 신호가 있다. 안전에 대한 확신을 더 얻기 위해 다음 훈련 실행을 3개월 연기해야 할까—하지만 안 된다, 최신 정보 보고서에 따르면 중국이 우리 모델 가중치를 훔쳐 자체 지능 폭발을 앞서가고 있다고 한다. 우리는 무엇을 해야 할까?”와 같은 미친 거래가 포함될 것이다.

나는 정부 프로젝트가 이 문제를 능숙하게 다룰 것이라고 확신하지 않는다—하지만 “스타트업이 개발한 초지능” 대안은 일반적으로 인식되는 것보다 훨씬 더 ‘최선을 기도하는’ 것에 가깝다. 우리는 이러한 어려운 거래를 수행하는 데 필요한 진지함을 테이블에 올릴 수 있는 지휘 체계가 필요할 것이다.

국제 상황 안정화

지능 폭발과 그 직후는 인류가 직면한 가장 불안정하고 긴장된 상황 중 하나를 불러올 것이다. 우리 세대는 이에 익숙하지 않다. 그러나 이 초기 기간에 당면한 과제는 멋진 제품을 만드는 것이 아니다. 어떻게든 필사적으로 이 시기를 견뎌내는 것이다.

우리는 정부 프로젝트가 경쟁에서 이기기를 필요로 할 것이다.

권위주의 세력에 맞서—그리고 이 상황의 위험을 헤쳐 나가기 위해 필요한 명확한 주도권과 숨 쉴 공간을 확보하기 위해서입니다. 만약 초지능 모델 가중치의 즉각적인 도난을 막지 못한다면, 우리는 포기하는 것이나 다름 없을 것입니다. 우리는 서방의 노력을 결집해야 합니다: 최고의 과학자들을 모으고, 찾을 수 있는 모든 GPU를 사용하며, 수조 달러 규모의 클러스터 구축이 미국 내에서 이루어지도록 해야 합니다. 우리는 데이터센터를 적의 사보타주나 직접적인 공격으로부터 보호해야 할 것입니다.

아마도 가장 중요한 것은 미국의 리더십일 것입니다—비확산 체제를 개발하고, 필요하다면 이를 강제하는 일입니다. 우리는 러시아, 북한, 이란, 그리고 테러 단체들이 자신들의 초지능을 이용해 세계를 인질로 잡을 수 있는 기술과 무기를 개발하지 못하도록 저지해야 합니다. 우리는 초지능을 활용해 우리의 핵심 인프라, 군사, 정부의 보안을 강화하여 극단적인 새로운 해킹 능력에 대비해야 합니다. 우리는 생물학이나 유사 분야의 발전에서 공격과 방어의 균형을 안정시키기 위해 초지능을 사용해야 합니다. 우리는 초지능을 안전하게 통제할 도구를 개발하고, 타인의 부주의한 프로젝트에서 나온 불량 초지능을 차단할 수 있어야 합니다. AI 시스템과 로봇은 인간 속도의 10~100배 이상으로 움직일 것이며, 모든 일이 매우 빠르게 진행될 것입니다. 우리는 한 세기 분량의 기술 진보를 몇 년 만에 압축하면서 발생할 수 있는 여섯 시그마 수준의 격변과 그에 따른 위협에 대비할 준비가 되어 있어야 합니다.

최소한 초기 기간 동안 우리는 가장 비상한 국가안보 위기에 직면할 것입니다. 아마도 이 임무를 수행할 사람이 없을지도 모릅니다. 그러나 우리가 가진 선택지 중에서, 프로젝트는 유일하게 합리적인 선택입니다.

프로젝트는 불가피하다; 그것이 좋은지는 알 수 없다

궁극적으로, 여기서 나의 주장은 기술적 설명에 가깝습니다: 우리가 원하든 원하지 않든, 초지능은 SF 스타트업처럼 보이지 않을 것이며, 어떤 면에서는 주로 국가안보 영역에 속할 것입니다. 지난 1년간 샌프란시스코 친구들에게 프로젝트에 대해 많이 이야기했는데, 아마도 가장 놀라웠던 점은

대부분 사람들이 이 아이디어에 얼마나 놀라는지에 관한 것이다. 그들은 단순히 그 가능성을 고려해본 적이 없다. 하지만 일단 고려해보면, 대부분은 이것이 명백해 보인다고 동의한다. 우리가 만들고 있다고 생각하는 것이 어느 정도 맞다면, 물론 결국 이것은 (어떤 형태로든) 정부 프로젝트가 될 것이다. 만약 어떤 연구소가 내일 문자 그대로 초지능을 개발한다면, 당연히 연방 정부가 개입할 것이다.

중요한 자유 변수 중 하나는 ‘여부’가 아니라 ‘언제’이다. 정부가 지능 폭발 한가운데에 있을 때까지 무슨 일이 일어나고 있는지 깨닫지 못할 것인가, 아니면 몇 년 전에 미리 알게 될 것인가? 정부 프로젝트가 불가피하다면, 일찍 시작하는 것이 더 낫다. 우리는 그 몇 년을 보안 긴급 프로그램을 수행하고, 주요 관계자들을 신속히 교육하고 준비시키며, 기능하는 통합 연구소를 구축하는 데 절실히 필요할 것이다. 정부가 마지막 순간에야 개입한다면 훨씬 더 혼란스러울 것이고 (비밀과 모델 가중치는 이미 도난당했을 것이다).

또 다른 중요한 자유 변수는 우리가 결집할 수 있는 국제 연합이다: 초지능 개발을 위한 더 긴밀한 민주주의 동맹과, 나머지 세계에 제공하는 더 광범위한 이익 공유 제안이다.

- 전자는 퀘벡 협정과 비슷할 수 있다: 처칠과 루스벨트가 핵무기 개발을 위해 자원을 공동으로 투입하고, 상호 동의 없이 서로 또는 타국에 사용하지 않기로 한 비밀 협약. 우리는 영국(Deepmind), 일본과 한국 같은 동아시아 동맹국(칩 공급망), 그리고 NATO/기타 핵심 민주주의 동맹국(더 넓은 산업 기반)을 참여시키고자 할 것이다. 통합된 노력은 더 많은 자원과 인재를 확보하고 전체 공급망을 통제하며, 안전, 국가안보, 군사적 도전에 대해 긴밀한 조정을 가능하게 하고, 초지능의 권력 행사를 견제하는 데 도움이 될 것이다.

- 후자는 ‘평화를 위한 원자력’, 국제원자력기구(IAEA), 핵확산금지조약(NPT)과 비슷할 수 있다. 우리는 초지능의 평화적 혜택을 더 넓은 국가 집단(비민주국가 포함)과 공유하겠다고 제안하고, 그들에 대해 공격적으로 초지능을 사용하지 않겠다고 약속해야 한다. 그 대가로 그들은 자체 초지능 프로젝트를 추구하지 않으며, ...

AI 시스템 배치에 대한 안전 약속을 하고, 이중 용도 적용에 대한 제한을 수용한다. 이 제안은 군비 경쟁과 확산에 대한 유인을 줄이고, 초지능 이후 세계 질서를 위한 미국 주도의 광범위한 연합을 형성하기를 바란다.

아마도 가장 중요한 자유 변수는 필연적인 정부 프로젝트가 능력 있을지 여부일 것이다. 어떻게 조직될 것인가? 어떻게 이 일을 해낼 수 있을까? 견제와 균형은 어떻게 작동하며, 합리적인 지휘 체계는 어떤 모습일까? 이에 대해 거의 관심이 기울여지지 않았다.¹¹⁹ 거의 모든 다른 AI 연구소와 AI 거버넌스 정치 활동은

결말

그래서 27/28년경에는 결말이 시작될 것이다. 28/29년경에는 지능 폭발이 진행 중일 것이며; 2030년까지 우리는 모든 힘과 위력을 갖춘 초지능을 소환할 것이다.

프로젝트 책임자로 누가 임명되든 엄청난 임무가 주어진다: 범용 인공지능(AGI)을 빠르게 구축하는 것; 수억 개의 GPU를 생산하기 위해 미국 경제를 전시 체제로 전환하는 것; 모든 것을 철저히 통제하고, 스파이를 제거하며, 중국공산당(CCP)의 전면 공격을 막아내는 것; 수억 대의 AGI가 AI 연구를 맹렬히 자동화하며, 1년에 10년 치 도약을 이루고 곧 가장 똑똑한 인간보다 훨씬 뛰어난 AI 시스템을 만들어내는 상황을 관리하는 것; 이 일이 탈선하여 인간 감독자를 제어하려는 악성 초지능을 만들어내지 않도록 상황을 유지하는 것; 그 초지능들을 활용해

119. 저의 진보 연구(Progress Studies) 동지들에게: 여러분이 이 문제를 깊이 고민해야 합니다. 이것이 여러분 지적 프로젝트의 정점이 될 것이기 때문입니다! 여러분은 미국 정부 연구 기관, 지난 반세기 동안의 쇠퇴, 다시 효과 있게 만들 방법을 연구하는 데 많은 시간을 쏟아왔습니다. 묻겠습니다: 우리는 어떻게 ‘프로젝트’를 효과적으로 만들 것입니까?

상황을 안정시키고 적보다 앞서 나가기 위해 필요한 신기술을 개발하고, 이를 통합해 미군을 신속히 재편하는 것; 이 모든 것을 아마도 역사상 가장 긴장된 국제 정세 속에서 조율하는 것. 그들은 정말 잘해야 할 것이다, 확실히 말하겠다.

이 여정에 함께하라는 부름을 받는 우리에게는 ... 스트레스가 클 것이다. 하지만 자유 진영을 위해 봉사하는 의무가 될 것이다.



그림 38:오펜하이머와 그로브스.

세계—그리고 전 인류. 우리가 이 시기를 무사히 넘기고 그 시절을 되돌아볼 수 있다면, 그것은 우리가 지금까지 해온 일 중 가장 중요한 일이 될 것이다. 그리고 그들이 찾게 될 어떤 안전한 시설이 오늘날의 터무니없이 과대 평가된 AI 연구자 생활의 쾌적함을 갖추고 있지는 않겠지만, 그렇게 나쁘지는 않을 것이다. SF는 이미 특이한 AI

연구자 대학가 같은 느낌이 들고; 아마도 이곳도 크게 다르지 않을 것이다. 낮에는 스케일링 곡선을 고민하며 땀 흘리고, 주말에는 모여서 AGI와 그날그날의 연구실 정치에 대해 수다 떠는, 똑같이 이상하게 작은 모임일 것이다.

단, 음—그 위험 부담은 너무나도 현실적일 것이다.

사막에서 보자, 친구들.



그림 39:시카고 대학교에서 첫 통제된 핵분열 반응 4주년을 기념하는 원자 과학자들의 재회.

V. 작별의 생각

만약 우리가 옳다면?

나는 1941년 봄을 오늘날까지도 기억한다. 그때 나는 핵폭탄이 단지 가능할 뿐만 아니라 불가피하다는 것을 깨달았다. 조만간 이러한 생각은 우리만의 특이한 것이 아닐 것이다. 곧 모든 사람이 그것에 대해 생각할 것이고, 어떤 나라가 그것을 실행에 옮길 것이다....

그리고 그것에 대해 이야기할 사람이 아무도 없었다. 나는 많은 밤을 잠 못 이루었다. 하지만 그것이 얼마나 매우 심각할 수 있는지 깨달았다. 그리고 그때부터 수면제를 복용하기 시작했다. 그것이 유일한 치료법이었고, 그 이후로 한 번도 끊지 않았다. 28년이 지났고, 그 28년 동안 단 한 번도 밤을 거르지 않은 것 같다. 제임스 채드위치 (물리학 노벨상 수상자이자 1941년 영국 정부의 핵폭탄 불가피성 보고서 저자, 이 보고서가 결국 맨해튼 프로젝트를 촉발시켰다)

10년이 끝나기 전에 우리는 초지능을 구축할 것이다. 이것이 이 시리즈 대부분의 주제였다. 샌프란시스코에서 만나는 대부분의 사람들에게 그 지점이 바로 화면이 꺼지는 순간이다. 하지만 그 다음 10년, 즉 2030년대는 적어도 그만큼이나 사건이 많을 것이다. 그 끝에는 세상이 완전히, 알아볼 수 없을 정도로 변모해 있을 것이다. 새로운 세계 질서가 형성될 것이다. 그러나 안타깝게도—그 이야기는 다음 기회로 미뤄야겠다.

지금은 마무리해야 한다. 마지막으로 몇 가지 말을 남기겠다.

범용 인공지능(AGI) 현실주의

이 모든 것은 깊이 숙고할 문제이며, 많은 이들이 그러지 못한다. “딥러닝이 한계에 부딪혔다!”라고 매년 외친다. 전문가들은 자신 있게 이것이 또 다른 기술 붐일 뿐이라고 말한다. 하지만 샌프란시스코 중심지에 있는 사람들 사이에서도 논의는 두 가지 근본적으로 진지하지 않은 구호로 양분되었다.

한쪽 끝에는 파멸론자들이 있다. 그들은 수년간 AGI에 집착해왔으며, 나는 그들의 선견지명에 많은 존경을 표한다. 그러나 그들의 사고는 경직되어 딥러닝의 경험적 현실과 동떨어져 있으며, 그들의 제안은 순진하고 실행 불가능하며, 매우 현실적인 권위주의 위협과 맞서지 못한다. 99%의 파멸 확률을 맹목적으로 주장하고 AI를 무기한 중단하자고 외치는 것은 분명히 올바른 길이 아니다.

반대편에는 e/acc(가속주의자)들이 있다. 좁게 보면 그들은 AI 발전이 계속되어야 한다는 좋은 점을 가지고 있다. 그러나 그들의 피상적인 트위터 허튼소리 뒤에는 속임수가 있다; AGI를 직시하기보다는 자신들의 포장 스타트업을 만들고 싶어하는 아마추어들이다. 그들은 미국 자유의 열렬한 수호자라고 주장하지만, 불미스러운 독재자들의 돈의 유혹을 뿌리치지 못한다. 사실 그들은 진짜 정체 정체론자들이다. 위협을 부인하려는 시도 속에서 AGI 자체를 부인하며, 결국 우리가 얻을 것은 멋진 채팅봇뿐이고, 그것들은 분명히 위험하지 않다고 주장한다. (내 기준으로는 매우 실망스러운 가속주의다.)

그러나 내가 보기에는, 이 분야에서 가장 똑똑한 사람들은 다른 관점, 즉 내가 AGI 현실주의라고 부를 제3의 길에 수렴했다. 핵심 원칙은 간단하다:

1. 초지능은 국가안보 문제다. 우리는 가장 똑똑한 인간보다 더 똑똑한 기계를 빠르게 만들고 있다. 이것은 또 다른 멋진 실리콘밸리 붐이 아니다; 이것은 무고한 오픈소스 소프트웨어 패키지를 작성하는 무작위 코더들의 커뮤니티가 아니다; 이것은 재미와 게임이 아니다. 초지능은 거칠 것이며, 인류가 만든 가장 강력한 무기가 될 것이다. 그리고 우리 중 누구에게든, 이것은 우리가 하는 일 중 가장 중요한 일이 될 것이다.
2. 미국이 선도해야 한다. 자유의 횃불은 시진핑이 범용 인공지능(AGI)을 먼저 획득하면 살아남지 못할 것이다. (그리고 현실적으로도, 미국의 리더십만이 안전한 AGI로 가는 유일한 길이다.) 이는 단순히 “일시 중지”할 수 없다는 것을 의미한다; 미국 내에서 AGI 클러스터를 구축하기 위해 미국의 전력 생산을 신속히 확장해야 한다는 뜻이다. 하지만 또한 아마추어 스타트업의 보안이 핵 비밀을 중국공산당(CCP)에 넘기는 식으로는 더 이상 안 된다는 의미이며, 핵심 AGI 인프라는 중동의 어떤 독재자가 아니라 미국이 통제해야 한다는 뜻이다. 미국 AI 연구소들은 국가 이익을 최우선으로 해야 한다.
3. 우리는 실수하지 말아야 한다. 초지능의 힘을 인식하는 것은 그 위험도 인식하는 것을 의미한다. 매우 현실적인 안전 위험이 존재하며; 이 모든 것이 잘못될 위험도 매우 현실적이다—인류가 상호 파멸을 초래할 파괴적 힘을 사용하기 때문이든, 아니면 우리가 소환하는 외계 종족이 아직 완전히 통제할 수 없는 존재이기 때문이든. 이 위험들은 관리 가능하지만—즉흥적으로 대처해서는 안 된다. 이러한 위험을 헤쳐 나가려면 지금까지 보여주지 못한 수준의 진지함을 가진 좋은 사람들이 필요하다.

가속이 심화됨에 따라, 논의는 더욱 격렬해질 것으로 예상된다. 하지만 내가 가장 바라는 것은 다가오는 무게를 느끼고 이를 엄숙한 의무의 부름으로 받아들이는 사람들이 있을 것이라는 점이다.

만약 우리가 옳다면?

이 시점에서 당신은 나와 다른 SF계 사람들이 완전히 미쳤다고 생각할지도 모른다. 하지만 잠시만 생각해 보라: 만약 그들이 옳다면? 이들은 이 기술을 발명하고 구축한 사람들이며; AGI가 이번 10년 내에 개발될 것이라고 생

각한다; 그리고 꽤 넓은 스펙트럼이 있지만, 그들 중 많은 이들이 내가 이 시리즈에서 설명한 대로 초지능으로 가는 길이 전개될 가능성을 매우 진지하게 받아들인다.

거의 확실히 나는 이야기의 중요한 부분을 잘못 이해했을 것이다; 만약 현실이 이 정도로 미친 방향으로 전개된다면, 오차 범위는 매우 클 것이다. 게다가, 처음에 말했듯이, 나는 가능성의 폭이 넓다고 생각한다. 하지만 나는 중요하다고 생각한다—

구체적이어야 한다. 그리고 이 연재에서 나는 현재 내가 믿는, 앞으로 남은 10년—즉 이번 10년 동안 가장 가능성이 높은 단일 시나리오를 제시했다.

왜냐하면—이것이 점점 현실로 다가오고 있기 때문이다, 매우 현실적으로. 몇 년 전만 해도, 적어도 나에게, 이런 생각들이 진지했지만—추상적이었고, 모델과 확률 추정 속에 격리되어 있었다. 이제는 매우 생생하게 느껴진다. 나는 그것을 볼 수 있다. AGI가 어떻게 만들어질지 볼 수 있다. 더 이상 인간 뇌 크기 추정이나 가설, 이론적 외삽에 관한 문제가 아니다—나는 기본적으로 AGI가 훈련될 클러스터와 언제 만들어질지, 우리가 사용할 대략적인 알고리즘 조합, 해결되지 않은 문제들과 그것들을 해결하는 경로, 중요한 사람들의 목록까지 말할 수 있다. 나는 그것을 볼 수 있다. 매우 생생하다. 물론, 2023년 초에 Nvidia에 전폭적으로 투자한 것은 훌륭했지만, 역사의 무게는 무겁다. 나는 이것을 선택하지 않았을 것이다.

하지만 가장 무서운 깨달음은 이 일을 처리할 특급 팀이 없다는 것이다. 어릴 적에는 세상이 영웅적인 과학자들, 뛰어난 군인들, 침착한 지도자들이 있어서 위기가 닥치면 그들이 구해줄 것이라는 영광스러운 관념이 있다. 현실은 그렇지 않다. 세상은 믿을 수 없을 만큼 작다; 겉모습이 벗겨지면, 보통 무대 뒤에 있는 몇몇 사람들이 실제 플레이어이며, 필사적으로 상황이 무너지지 않도록 애쓰고 있다.

지금 이 순간, 아마도 전 세계에 몇 백 명 정도만이 다가올 충격을 인지하고, 상황이 얼마나 미쳐 돌아갈지 이해하며, 상황 인식을 갖고 있다. 나는 아마도 프로젝트를 실행할 수 있는 모든 사람을 개인적으로 알거나 한 단계 정도의 인맥 거리에 있다. 무너지는 것을 막으려 필사적으로 애쓰는 무대 뒤의 몇몇 사람들은 바로 당신과 당신의 친구들, 그리고 그들의 친구들이다. 그게 전부다. 그것이 전부다.

언젠가는 이 일이 우리 손을 벗어날 것이다. 하지만 적어도 앞으로 몇 년간 중반 게임 동안은, 세계의 운명이 이 사람들에게 달려 있다.

자유 진영이 승리할 것인가?

우리는 초지능을 길들일 것인가, 아니면 초지능이 우리를 길들일 것인가?

인류는 또다시 자기 파괴를 피할 수 있을 것인가?

위험은 결코 줄어들지 않았다.

이들은 위대하고 존경받을 만한 사람들이지만, 결국 한낱 사람들일 뿐이다. 곧 AI가 세상을 지배하게 될 것이지만, 우리는 마지막 한 번의 결전을 치러야 한다. 그들의 마지막 관리가 인류에게 영광을 가져다주길 바란다.

부록

컴퓨팅 대략 계산에 대한 추가 세부사항

훈련 컴퓨팅

연도. OpenAI GPT-4 기술 보고서에 따르면 GPT-4는 2022년 8월에 훈련을 마쳤다. 이후 우리는 대략 연간 0.5 규모 차수(OOM)의 추세를 적용한다.

H100s-동등. Semianalysis, JP Morgan 등은 GPT-4가 25,000대의 A100으로 훈련되었다고 추정하며, H100은 A100 대비 2~3배의 성능을 가진다.

비용. 종종 사람들은 “GPT-4 훈련에 1억 달러” 같은 숫자를 GPU 임대 비용만으로 인용한다(즉, “이 정도 규모의 클러스터를 3개월간 임대하는 데 드는 비용” 같은 것). 하지만 이는 잘못된 접근이다. 중요한 것은 클러스터를 실제로 구축하는 데 드는 비용에 가깝다. 세계 최대 규모 클러스터 중 하나를 원한다면 단지 3개월 임대하는 것으로는 부족하다! 게다가 컴퓨팅은 단지 주력 훈련 실행뿐 아니라 위험 완화 실험, 실패한 실행, 다른 모델 등에도 필요하다.

GPT-4 클러스터 비용을 대략 추산하기 위해:

- 공개된 추정치는 GPT-4 클러스터가 약 25,000대의 A100으로 구성되어 있음을 시사한다.
- 2-3년 동안 \$1/A100-시간을 가정하면 대략 \$5억의 비용이 듭니다.
- 또는 H100당 \$25,000 비용, 10,000 H100 상당, 그리고 Nvidia GPU가 클러스터 비용의 약 절반(나머지는 전력, 물리적 데이터센터, 냉각, 네트워킹, 유지보수 인력 등)이라는 점을 고려해 추정할 수 있습니다. (예를 들어, 이 총소유비용 분석에 따르면 대형 클러스터 비용의 약 40%가 H100 GPU 자체에 해당하며, 추가로 13%는 Infiniband 네트워킹을 위한 Nvidia 비용입니다. 다만, 자본 비용을 제외하면 GPU가 비용의 약 50%를 차지하고, 네트워킹 비용까지 포함하면 Nvidia가 클러스터 비용의 60% 이상을 차지합니다.)

FLOP/\$는 각 Nvidia 세대마다 다소 개선되고 있지만 큰 폭은 아닙니다. 예를 들어, H100 -> B100은 약 1.5배 FLOP/\$ 개선일 가능성이 있습니다: B100은 사실상 두 개의 H100을 결합한 것이지만, 가격은 2배 미만입니다. 그러나 B100은 다소 예외적인 경우로, Nvidia가 경쟁을 압도하려는 의도로 가격을 낮춘 것으로 보입니다. 반면 A100 -> H100은 FLOP/\$ 개선이 크지 않았습니다(비FP8 기준으로 2배 성능 향상에 비용도 약 2배, FP8 개선을 포함하면 약 1.5배) — 그리고 이는 2년 주기의 세대였습니다.

마진 압박으로 인해 FLOP/\$가 더 개선될 여지가 있다고 생각하지만, GPU가 대규모 제약을 받으면서 오히려 더 비싸질 수도 있습니다. AI 칩 특화로 인한 이득은 계속되겠지만, 칩이 이미 AI에 꽤 특화되어 있고(예: 트랜스포머에 특화, 이미 fp8/fp4 정밀도 사용), 무어의 법칙은 현재 매우 느리며, 메모리와 인터커넥트 같은 병목 부품의 개선 속도도 더딘 점을 고려하면 FLOP/\$에 대한 획기적인 기술적 개선이 있을지는 불확실합니다. Epoch의 데이터를 보면 지난 10년간 최고 ML GPU의 FLOP/\$는 10배 미만의 개선을 보였고, 앞서 언급한 이유들로 인해 앞으로는 더 느려질 것으로 예상됩니다.

연간 약 35%의 FLOP/\$ 개선이 있다면 +4 OOM 클러스터에 1조 달러 비용이 들 것입니다. FLOP/\$가 더 빠르게 개선될 수도 있지만, 데이터센터 자본 지출도 더 늘어날 것입니다.

비용이 많이 든다—단순히 기존 감가상각된 발전소를 임대하는 대신 실제로 새로운 전력을 구축해야 하므로 초기 자본 지출(capex)이 많이 필요하다.

어쨌든 이 숫자들은 매우 대략적인 것이다. 예를 들어 1T 클러스터가 더 효율적으로 구축되어 실제로 컴퓨팅에서 +4.5 규모 차수(OOM) 정도의 이득을 낸다면 오차 범위 내에 충분히 들어갈 수 있다.

비용. H100은 700W이지만, 데이터센터 전력(냉각, 네트워킹, 저장소 등)도 필요하다; Semianalysis는 H100 당 약 1,400W로 추정한다.

FLOP/와트에서 얻을 수 있는 이득이 있긴 하지만, AI 칩 특화에서 얻을 수 있는 이득이 다 소진된 후(이전 각주 참조), 예를 들어 가장 낮은 정밀도로 내려간 후에는 이러한 이득이 다소 제한적이다(주로 칩 공정 개선인데, 이는 느리다). 그렇긴 해도 전력이 더 큰 제약이 되고(따라서 비용에서 차지하는 비중이 커지면) 칩 설계가 FLOP 비용을 희생하면서 더 전력 효율적으로 특화될 수도 있다. 그럼에도 불구하고 냉각, 네트워킹, 저장소 등에 대한 전력 수요는 여전히 존재한다(위 H100 수치에서는 이미 전력 수요의 약 절반 정도였다).

이 대략적인 계산을 위해 여기서는 H100 등가당 1kW를 사용하자; 다시 말하지만, 이들은 단지 대략적인 계산이다. (만약 예상치 못한 FLOP/와트 혁신이 있다면, 동일한 전력 소비를 유지하면서 더 큰 규모 차수 컴퓨팅 이득을 기대할 수 있을 것이다.)

전력 참조 기준. 10GW 클러스터를 1년 내내 가동하면 87.6 TWh가 소모된다. 비교하자면, 오리건 주는 연간 약 27 TWh의 전기를 소비하고, 워싱턴 주는 연간 약 92 TWh를 소비한다.

100GW 클러스터를 1년 내내 가동하면 876 TWh가 소모되며, 미국 전체 연간 전기 생산량은 약 4,250 TWh이다.

여기를 클릭하여 “학습 컴퓨팅” 표로 돌아가기.

전체 컴퓨팅

올해 H100 동등량. 저는 Nvidia가 2024년에 약 500만 개의 데이터센터 GPU를 출하할 것으로 추정합니다. 그 중 소수는 B100으로, 이를 H100의 2배 이상으로 계산할 것입니다. 그리고 다른 AI 칩들도 있습니다: TPU, Trainium, Meta의 맞춤형 실리콘, AMD GPU 등.

TSMC 용량. TSMC는 월 15만 개 이상의 5nm 웨이퍼 생산 능력을 보유하고 있으며, 월 10만 개의 3nm 웨이퍼로 생산을 확대 중이고, 아마도 월 15만 개 정도의 7nm 웨이퍼도 생산할 것입니다. 총합으로 월 약 40만 개의 웨이퍼 생산 능력이 있다고 할 수 있습니다.

웨이퍼당 대략 35개의 H100이 생산된다고 가정합니다(H100은 5nm 공정으로 제조됨). 2024년에 500만 ~1000만 개의 H100 동등량을 생산한다면, 연간 AI 칩 생산량은 15만~30만 개의 웨이퍼에 해당합니다.

이 범위 내 어디에 위치하는지, 그리고 7nm 생산을 포함할지 여부에 따라, 이는 연간 최첨단 웨이퍼 생산량의 약 3~10%에 해당합니다.

여기를 클릭하여 “전체 컴퓨팅” 표로 돌아가세요.

미주

본문의 위첨자 번호는 원문에 포함된 번호형 주석을 가리킵니다. 이미지 설명 캡션은 원래 위치에 유지했고, 본문 설명 주석만 이 미주 영역으로 분리했습니다.