TIME SUBSCRIBE

TECH ARTIFICIAL INTELLIGENCE

4 Charts That Show Why AI Progress Is Unlikely to Slow Down

9 MINUTE READ

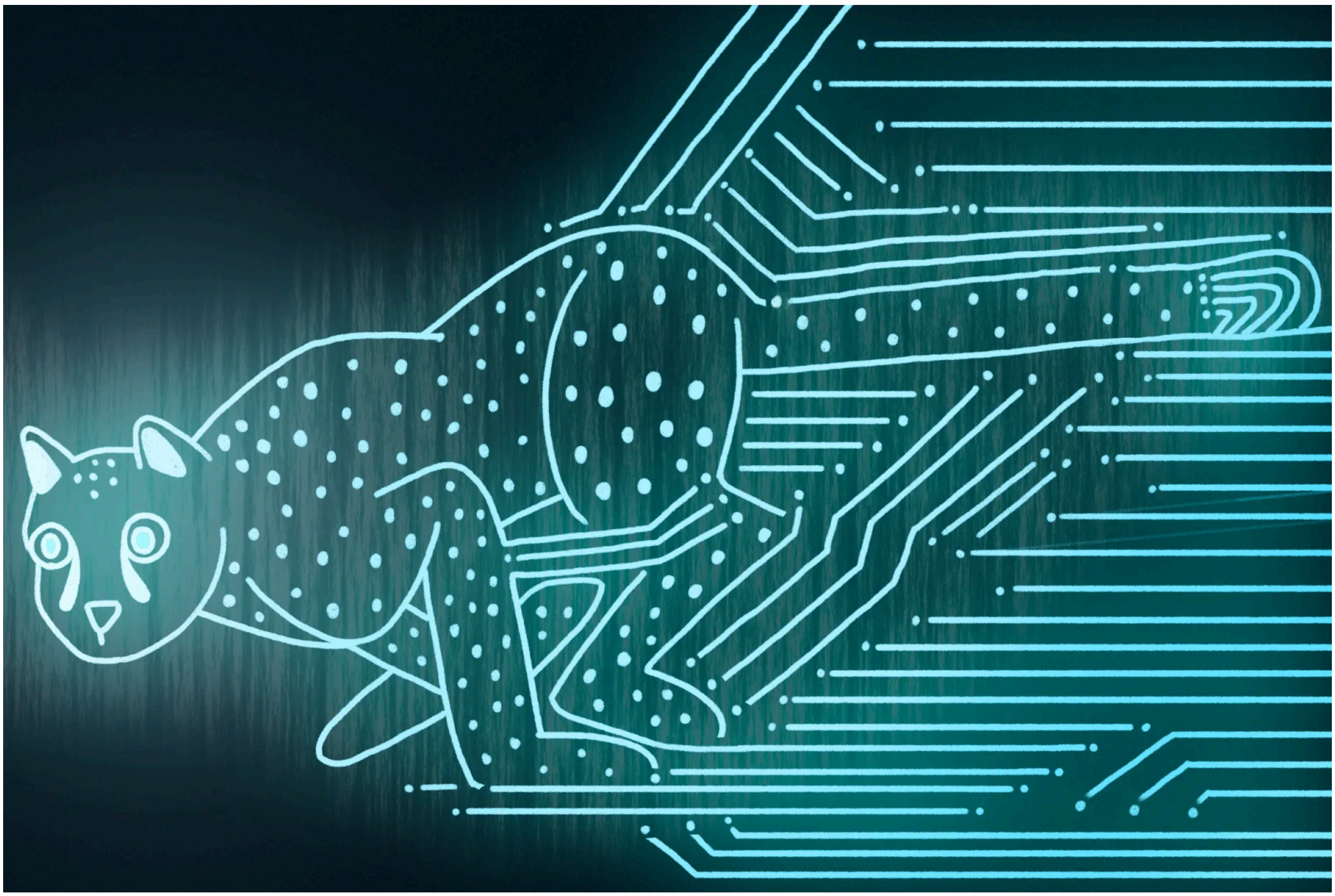


Illustration by Katie Kalupson for TIME

BY **WILL HENSHALL**

UPDATED: NOVEMBER 6, 2023 4:13 PM EST | ORIGINALLY PUBLISHED: AUGUST 2, 2023 4:50 PM EDT

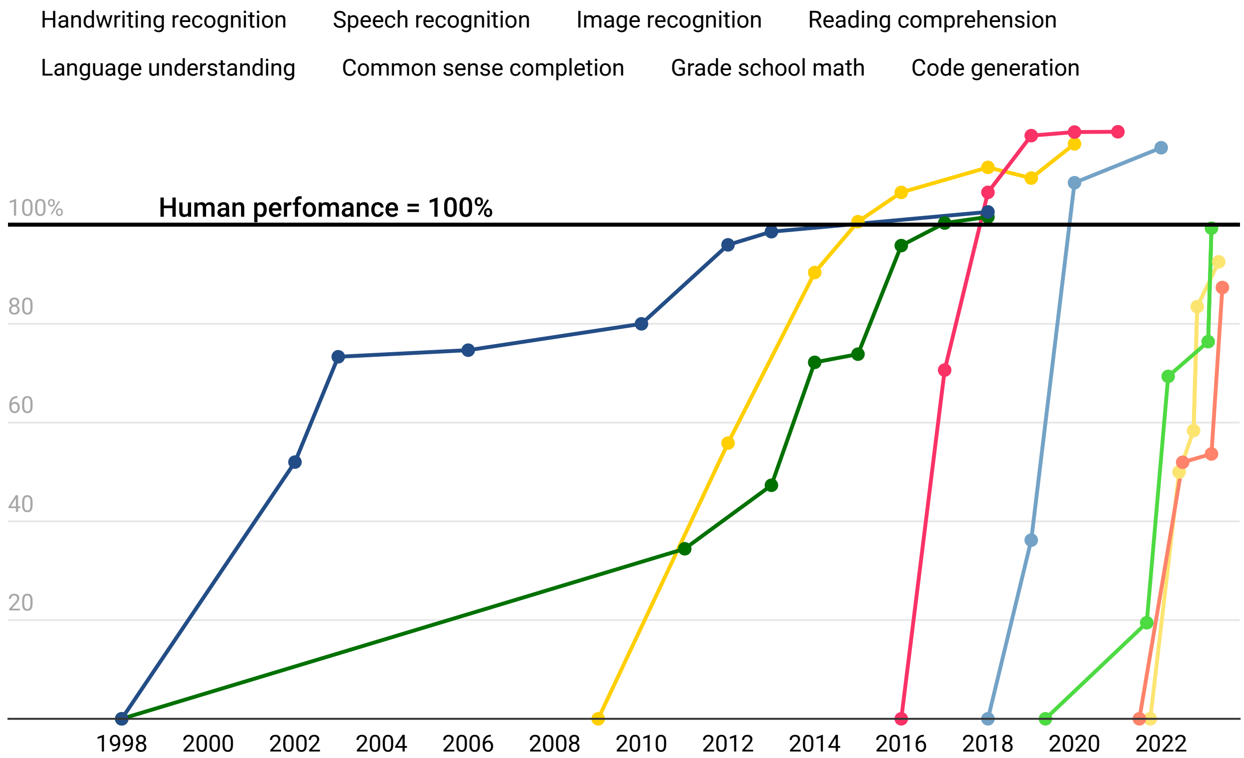
In the last ten years, AI systems have developed at rapid speed. From the breakthrough of besting a legendary player at the complex game Go in 2016, AI is now able to recognize images and speech better than humans, and pass tests including business school exams and Amazon coding interview questions.

Last week, during a U.S. Senate Judiciary Committee hearing about regulating AI, Senator Richard Blumenthal of Connecticut described the reaction of his constituents to recent advances in AI. “The word that has been used repeatedly is scary.”

The Subcommittee on Privacy, Technology, and the Law overseeing the meeting heard testimonies from three expert witnesses, who stressed the pace of progress in AI. One of those witnesses, Dario Amodei, CEO of prominent AI company Anthropic, said that “the single most important thing to understand about AI is how fast it is moving.”

AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

State-of-the-art AI performance on benchmarks, relative to human performance



For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the "starting point", which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwag, Code generation = HumanEval.

Chart: Will Henshall for TIME • Source: ContextualAI TIME

It's often thought that scientific and technological progress is fundamentally unpredictable, and is driven by flashes of insight that are clearer in hindsight. But progress in the capabilities of AI systems is predictably driven by progress in three inputs—compute, data, and algorithms. Much of the progress of the last 70 years has been a result of researchers training their AI systems using greater computational processing power, often referred to as "compute", feeding the systems more data, or coming up with algorithmic hacks that effectively decrease the amount of compute or data needed to get the same results. Understanding how these three factors have driven AI progress in the past is key to understanding why most people working in AI don't expect progress to slow down any time soon.

Read more: *The AI Arms Race Is Changing Everything*

Compute

The first artificial neural network, Perceptron Mark I, was developed in 1957 and could learn to tell whether a card was marked on the left side or the right. It had 1,000 artificial neurons, and training it required around 700,000 operations. More than 65 years later, OpenAI released the large language model GPT-4. Training GPT-4 required an estimated 21 septillion operations.

Increasing computation allows AI systems to ingest greater amounts of data, meaning the system has more examples to learn from. More computation also allows the system to model the relationship between the variables in the data in greater detail, meaning it can draw more accurate and nuanced conclusions from the examples it is shown.

More From TIME

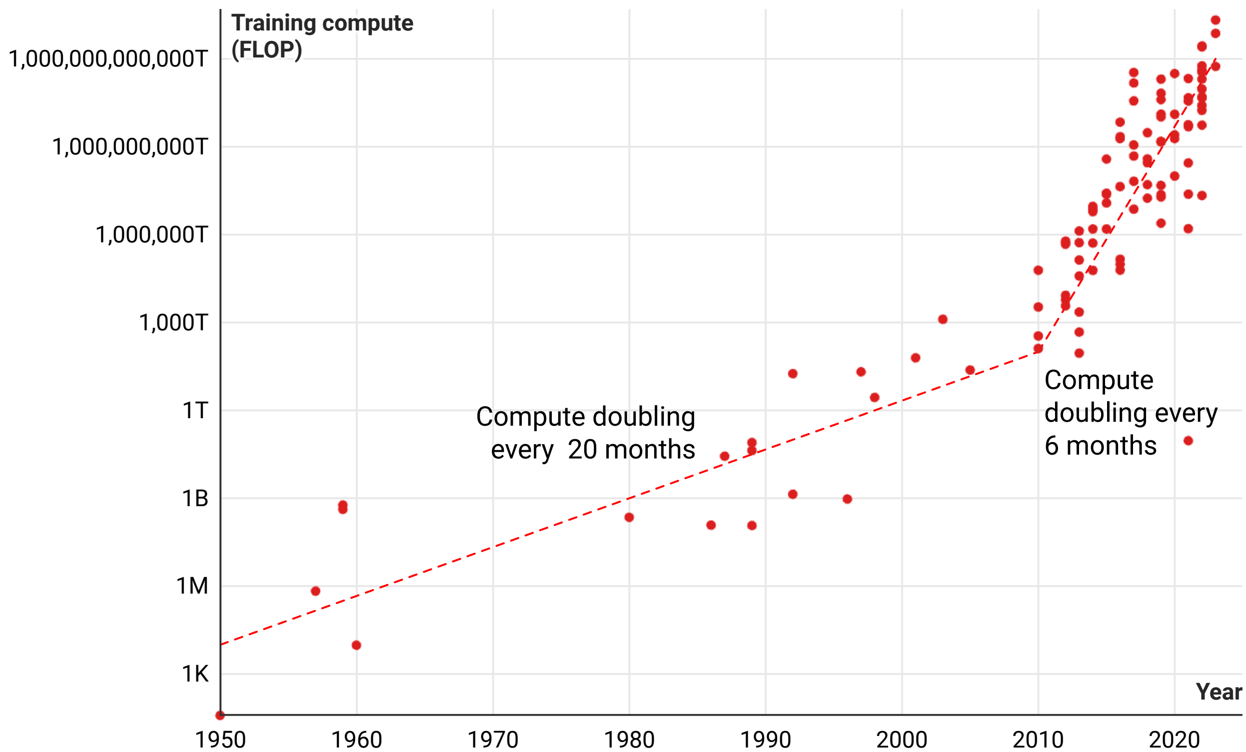
Since 1965, Moore's law—the observation that the number of transistors in an integrated circuit doubles about every two years—has meant the price of compute has been steadily decreasing. While this did mean that the amount of compute used to train AI systems increased, researchers were more focused on developing new techniques for building AI systems rather than focusing on how much compute was used to train those systems, according to Jaime Sevilla, director of Epoch, a research organization.

This changed around 2010, says Sevilla. "People realized that if you were to train bigger models, you will actually not get diminishing returns," which was the commonly held view at the time.

Since then, developers have been spending increasingly large amounts of money to train larger scale models. Training AI systems requires expensive specialized chips. AI developers either build their own computing infrastructure, or pay cloud computing providers for access to theirs. Sam Altman, CEO of OpenAI, has said that GPT-4 cost over $100 million to train. This increased spending, combined with the continued decreases in the cost of the increases in compute resulting from Moore's Law, has led to AI models being trained on huge amounts of compute.
OpenAI and Anthropic, two of the leading AI companies, have each raised billions from investors to pay for the compute they use to train AI systems, and each has partnerships with tech giants that have deep pockets—OpenAI with Microsoft and Anthropic with Google.

The amount of compute used to train AI systems has been increasing since 1950, the rate of increase increased in 2010

Amount of compute used to train notable AI models



FLOP (floating-point operations) refers to the total number of computer operations used to train an AI system. Computation is estimated based on published results in the AI literature and comes with some uncertainty. Epoch expect most of these estimates to be correct within a factor of 2, and a factor of 5 for recent models for which relevant numbers were not disclosed, such as GPT-4.

Chart: Will Henshall for TIME • Source: Epoch via Our World in Data

TIME

Data

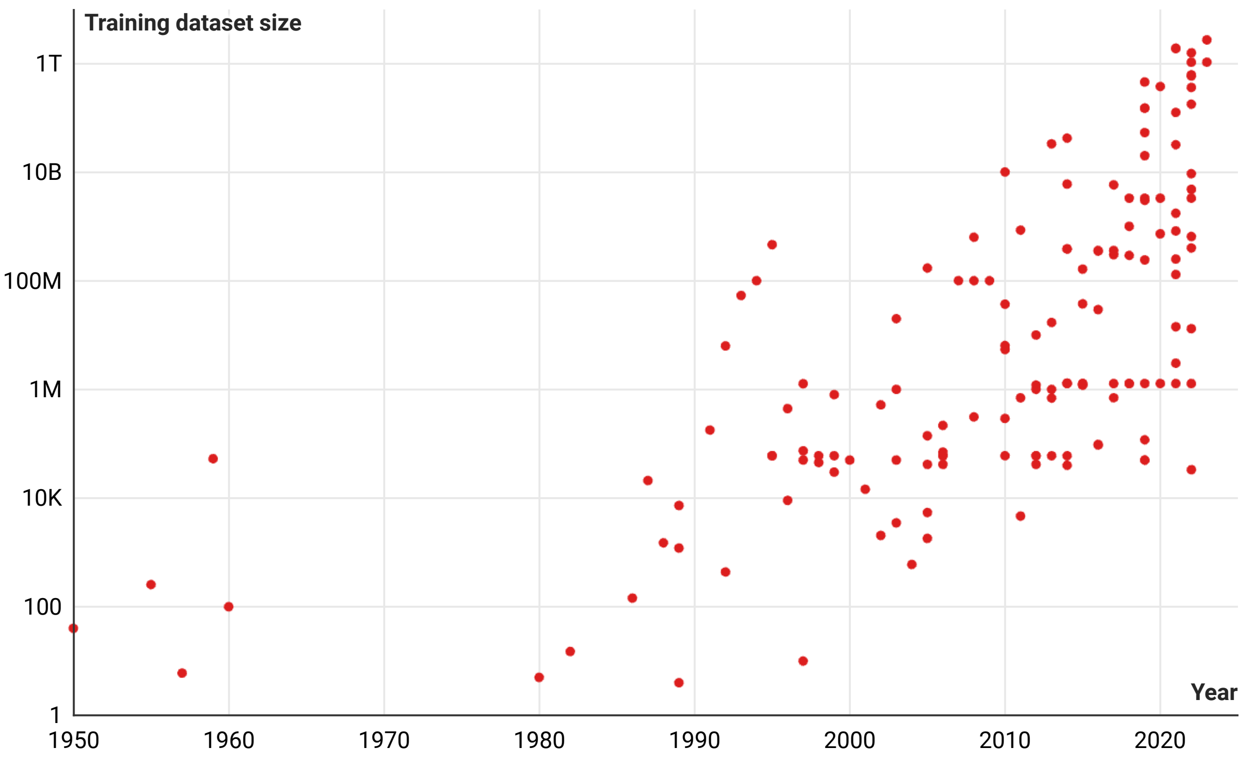
AI systems work by building models of the relationships between variables in their training data—whether it's how likely the word "home" is to appear next to the word "run," or patterns in how gene sequence relates to protein folding, the process by which a protein takes its 3D form, which then defines its function.

In general, a larger number of data points means that AI systems have more information with which to build an accurate model of the relationship between the variables in the data, which improves performance. For example, a language model that is fed more text will have a greater number of examples of sentences in which the "run" follows "home"—in sentences that describe baseball games or emphatic success, this sequence of words is more likely.

The original research paper about Perceptron Mark I says that it was trained on just six data points. By comparison, LlaMa, a large language model developed by researchers at Meta and released in 2023, was trained on around one billion data points—a more than 160-million fold increase from Perceptron Mark 1. In the case of LlaMa, the data points was text collected from a range of sources, including 67% from Common Crawl data (Common Crawl is a non-profit that scrapes the internet and makes the data collected freely available), 4.5% from GitHub (an internet service used by software developers), and 4.5% from Wikipedia.

The number of data points used to train AI models has increased dramatically over the last seventy years

Number of data points used to train notable AI models



Training data size refers to the amount or quantity of data that is used to train an AI model, indicating the number of examples or instances available for the model to learn from. Each domain has a specific data point input unit, such as images to train vision models, words for language models and timesteps for games models. This means systems can only be compared directly within the same domain.

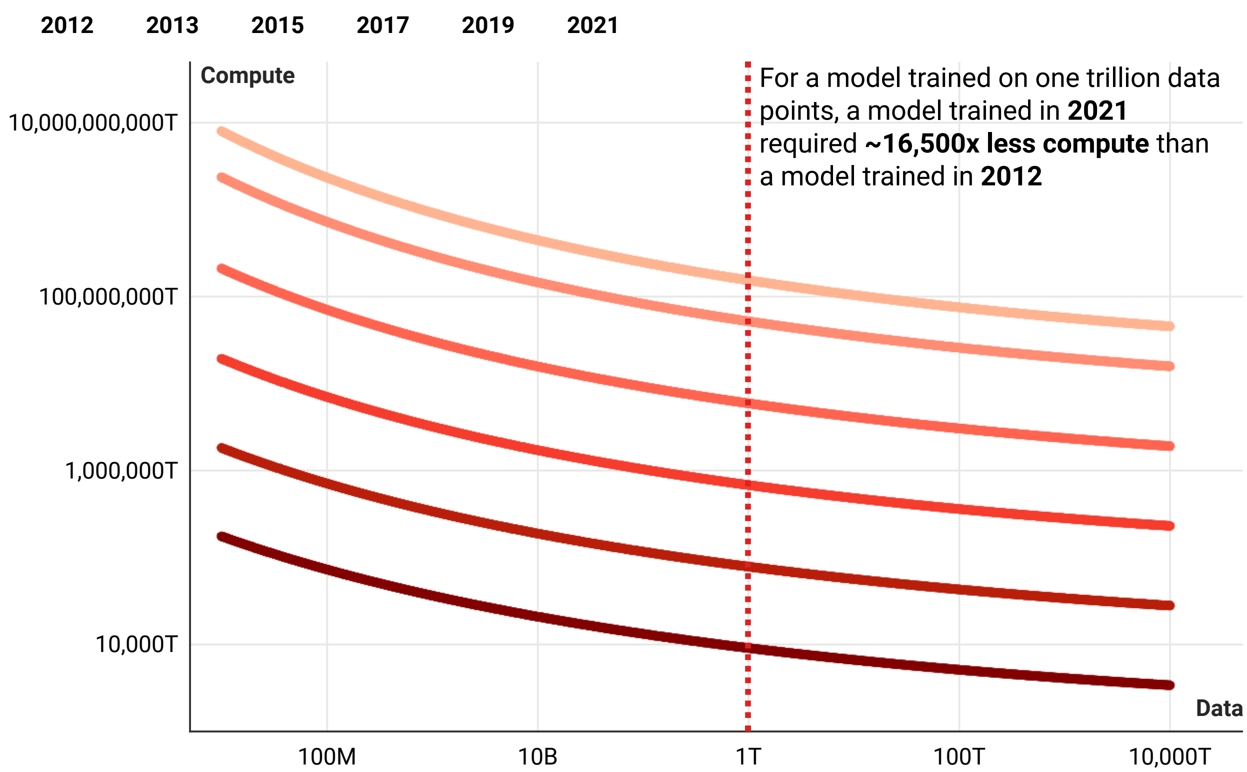
Chart: Will Henshall for TIME • Source: Epoch via Our World in Data TIME

Algorithms

Algorithms—sets of rules or instructions that define a sequence of operations to be carried out— determine how exactly AI systems use computational horsepower to model the relationships between variables in the data they are given. In addition to simply training AI systems on greater amounts of data using increasing amounts of compute, AI developers have been finding ways to get more from less. Research from Epoch found that "every nine months, the introduction of better algorithms contributes the equivalent of a doubling of computation budgets."

Algorithmic progress means that less compute and data are required to achieve a given level of performance

Amount of compute and number of data points required to achieve 80.9% accuracy on an image recognition test



ResNeXt-101 computer vision system on the ImageNet benchmark. Compute is measured in FLOPs (floating-point operations). Data is measured in the number of images in the training set.

Chart: Will Henshall for TIME • Source: Epoch

TIME

The next phase of AI progress

According to Sevilla, the amount of compute that AI developers use to train their systems is likely to continue increasing at its current accelerated rate for a while, with companies increasing the amount of money they spend on each AI system they train, and with increased efficiency as the price of compute continues to decrease steadily. Sevilla predicts that this will continue until at some point it is no longer worth it to keep spending more money, when increasing the amount of compute only slightly improves performance. After that, the amount of compute used will continue to increase, but at a slower rate solely due to the cost of compute decreasing as a result of Moore's law.

The data that feeds into modern AI systems, such as LlaMa, is scraped from the internet. Historically, the factor limiting how much data is fed into AI systems has been having enough compute to process that data. But, the recent explosion in the amount of data used to train AI systems has outpaced the production of new text data on the internet has led researchers at Epoch to predict that AI developers will run out of high-quality language data by 2026.

Those developing AI systems tend to be less concerned about this issue. Appearing on the Lunar Society podcast in March, Ilya Sutskever, chief scientist at OpenAI, said that "the data situation is still quite good. There's still lots to go." Appearing on the Hard Fork podcast in July, Dario Amodei estimated that "there's maybe a 10% chance that this scaling gets interrupted by inability to gather enough data."

Sevilla is also confident that a dearth of data won't prevent further AI improvements—for example by finding ways to use low-quality language data—because unlike compute, lack of data hasn't been a bottleneck to AI progress before. He expects there to be lots of low hanging fruit in terms of innovation that AI developers will likely discover to address this problem.

Algorithmic progress, Sevilla says, is likely to continue to act as an augmenter of how much compute and data is used to train AI systems. So far, most improvements have come from using compute more efficiently. Epoch found that more than three quarters of algorithmic progress in the past has been used to make up for shortfalls in compute. If in future, as data becomes a bottleneck for progress on AI training, more of the algorithmic progress may be focused on making up for shortfalls in data.

Putting the three pieces together, experts including Sevilla expect AI progress to continue at breakneck speed for at least the next few years. Compute will continue to increase as companies spend more money and the underlying technology becomes cheaper. The remaining useful data on the internet will be used to train AI models, and researchers will continue to find ways to train and run AI systems which make more efficient use of compute and data. The continuation of these decadal trends is why experts think AI will continue to become more capable.

This has many experts worried. Speaking at the Senate Committee hearing, Amodei said that, if progress continues at the same rate, a wide range of people could be able to access scientific know-how that even experts today do not have within the next two to three years by using AI systems. This could increase the number of people who can "wreak havoc," he said. "In particular, I am concerned that AI systems could be misused on a grand scale in the domains of cybersecurity, nuclear technology, chemistry, and especially biology."

Correction, Nov. 6:

The original version of this story stated that GPT-4 was released more than 70 years after Perceptron Mark I was developed. It was released 66 years after.

WRITE TO WILL HENSHALL AT WILL.HENSHALL@TIME.COM

No Sympathy From Family for Some Axed Federal Workers

Some laid-off federal workers are running into another unexpected unpleasantry: Relatives cheering their firing.

TIME

What We Know About the Venezuelans Deported to El Salvador

Many of the men, including Jerce Barrios, have no criminal records, and some gang membership allegations are falling apart.

TIME

'He's Going to Tank Our Economy': Congress Reacts to Trump Tariffs

Trump's tariffs are fueling the strongest GOP pushback since Trump's inauguration. Democrats say Trump is wrecking the economy.

TIME

TIME