MERCATUS SPECIAL STUDY

COMPOUNDING INTELLIGENCE
ADAPTING TO THE AI REVOLUTION

Nabeel S. Qureshi

MERCATUS.ORG

MERCATUS CENTER
George Mason University

Nabeel S. Qureshi, "Compounding Intelligence: Adapting to the AI Revolution," Mercatus Special Study, Mercatus Center at George Mason University, Arlington, VA, July 2024.

ABSTRACT

In recent years and months, artificial intelligence (AI) has been catching up to human performance at complex cognitive tasks. This paper examines the drivers of this trend—increasing computational power, expanding data availability, and improving algorithmic efficiency—and argues that the trend is likely to continue. The paper then reviews the implications of this trend for national security, policy, and the economy. AI is likely to be one of the most important productivity-enhancing innovations of our time, and key policy decisions about AI made now will resonate through the coming decades. The paper outlines a set of policies that would promote economic growth and enable AI to be deployed to enhance dynamism in the American economy, while arguing that enhanced societal resilience, better evaluations, and monitoring frameworks are sufficient to address risks. The paper advocates for an empirical, evidence-based approach to AI policy and regulation, and it cautions against premature and overly restrictive measures that could stifle innovation.

JEL codes: O30, O33, O38, L86, K24

Keywords: AI, artificial intelligence, AGI, artificial general intelligence, machine learning, large language models, compute scaling, data scaling, AI policy, AI regulation, AI safety, AI ethics, technological innovation, AI evaluation, open source AI, AI governance, algorithmic efficiency, AI risk assessment, synthetic data, AI economic impact

Some sections of this study were originally published as essays on the *Digital Spirits* Substack (https://digitalspirits.substack.com/).

COMPOUNDING INTELLIGENCE

In 2009, the computer scientist Shane Legg—who went on to co-found DeepMind, an artificial intelligence (AI) research lab—predicted that artificial general intelligence (AGI) would match human capabilities between 2025 and 2028.[1]

As of 2024, OpenAI's GPT-4 seems to approach a human performance level on many tasks and exceed it on some. If AI continues to develop as fast as this trendline suggests, the world will look radically different by 2030, and nations will need to adapt. COVID-19's rapid spread caught many world governments flat-footed; AI's phenomenal growth could do the same.

Legg's 2009 insight was that intelligence is a function of compute (capacity for performing computational tasks), and a crucial driver of compute is Moore's law.[2] Decades ago, Intel co-founder Gordon Moore observed that the number of transistors in an integrated circuit had doubled every two years and projected this trend would continue for at least another decade. This assertion has had astonishing longevity. Combining his insight on intelligence with Moore's Law, Legg was able to extrapolate to the AI of the present day.

This study provides an analytical introduction to the AI revolution now taking place. It is intended to ground a nontechnical audience in some of the major trends and draw preliminary conclusions on policy, regulation, and key investments that could help ensure that the AI revolution goes as well as possible.

Predictable growth?

An under-appreciated aspect of Moore's law is that the improvements Moore predicted were not from a single source. Instead, a constant stream of novel improvements seemed to follow Moore's prediction.

---

1. Shane Legg, "Tick, Tock, Tick, Tock . . . ," *Vetta Project* (blog), February 9, 2009, https://www.vetta.org/2009/12/tick-tock-tick-tock-bing/.
2. Gordon E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics* 38, no. 8 (April 19, 1965): 114–17.

Between 1965 and 2005, Dennard scaling[3] (ever-smaller transistors resulting in more transistors per chip) was the primary driver of the technology curve Moore's law describes. Eventually, though, Dennard scaling began to reach its physical limits and break down, leading many to conclude that Moore's law was petering out.

But unexpectedly, Moore's law continued to prove true. From 2005 to 2020, growth in the number of transistors per chip was primarily driven by innovations like die size increases and bigger chips. New innovations like nanomaterials are poised to drive another decade of growth.

Moore's law was not just a prediction or extrapolation. It also functioned as an aspirational goal for the semiconductor industry, motivating it to maintain a high rate of innovation to keep up with the curve.

The same is true for modern AI: we can expect improvements to continue and compound over time.

To understand why, we need to understand the factors that go into improving AI models. We can talk about four major factors:

1. **Quantity and quality of data.** Large language models (LLMs) are trained on enormous bodies of data from the internet and digitized books, among other sources. Larger, higher-quality datasets yield bigger gains in model performance.
2. **Compute and scale of model.** The scale of an LLM is measured by the number of its parameters: the larger the model, the more parameters. The number of parameters, in turn, correlates to the amount of compute needed to train the model. This boils down to the hardware resources—central processing units (CPUs) and graphics processing units (GPUs)—used to train the model (i.e., compute capacity). Compute capacity is doubling every 9 to 10 months.
3. **Algorithmic/sample efficiency.** There are compelling reasons to believe that models have been learning inefficiently. After all, infants do not need to ingest billions of words from the internet to learn how to speak! Indeed, new techniques are enabling LLMs to learn far more efficiently from samples, which in turn reduces the amount of data required to achieve a given level of AI capability.

3. R. Dennard et al., "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE Journal of Solid-State Circuits* 9, no. 5 (October 1974): 256–68.

4. **Other innovations.** Several other innovations have had a dramatic effect on AI performance. For example, the invention of the transformer took LLMs across a critical threshold to where they were usable and conversational agents such as ChatGPT, and reinforcement learning from human feedback (RLHF) makes GPT-3.5 feel more lifelike—and perhaps closer to passing the Turing test—than GPT-3. Another example might be true synthetic data generation for LLMs (see essay 3 for more discussion of this).

These factors, especially data and compute, are quantifiable, and they grow at predictable rates. The sections that follow will dive deeper into both factors. For now, it's worth noting that future LLM capabilities can be extrapolated with a surprising degree of accuracy.

Implications

If this analysis of AI improvement based on its factors of production is correct, it has some important, counter-intuitive implications.

First, human-level AI will eventually be trainable by someone like a graduate student using the resources offered in any standard-sized university lab cluster. Although current AI models can be trained only with billions of dollars in capital, the cost and efficiency curves at play suggest that in the future, powerful AI models will not need to be especially large.

For example, when ChatGPT was launched, OpenAI founder and CEO Sam Altman estimated that the product cost "several cents" to run per chat.[4] As of publication, services of the open-source model Mixtral 8x7b are being offered at 50 cents per million tokens, which is one-hundredth of the cost from a year earlier. We can expect that, with more powerful models like GPT-4, their costs will continue to decrease rapidly after launch.

Second, proprietary models will eventually outpace open-source ones, unless a large tech company makes the weights for its powerful models freely available. More powerful LLMs than the GPT-4 generation will require exponentially greater costs and proprietary data to train, so only the largest tech companies will be able to afford them. Chip improvements and algorithmic efficiency will help to lower this hurdle, but only partway. Open-source models will be able to mimic these larger models and achieve some fraction of

4. Sam Altman (@sama), "average is probably single-digits cents per chat; trying to figure out more precisely and also how we can optimize it," Twitter (now X), December 5, 2022, 2:46 a.m., https://x.com/sama/status/1599671496636780546.

their performance; however, investing in data and compute for open-source developers and the academic community will be important to help them keep up.

Third, attempts to curb AI development would be futile, because the trends driving its growth would continue anyway. Even if governments attempted to regulate or restrain AI development, chips would continue to get faster, and open-source models would still improve. As a result, AGI would remain attainable for any capable academic department with access to a lab cluster, not to mention foreign countries such as China. Any measures taken in the United States to slow down AI model development would make it less prepared for a world where all its rivals have AGI.

The best way forward with AI is empirical and pragmatic: develop and test models, observe what happens, and learn as we go.

The returns on intelligence

The surprising predictability of the data- and compute-scaling curves puts us in the unusual situation of being able to forecast to reasonable resolution when human-level AI might be achieved on various dimensions. Though the precise definition of true human-level AGI is unclear, its arrival will be a matter of degree rather than a sharp distinction.

The biggest unknown, with the largest implications if true, is whether or when AI will become self-improving, able to contribute to research on itself. This is already happening in a sense; AI assistants write a percentage of code today. But the key question is whether improvements in AI will then lead to AI R&D improvement that is disproportionately large (accelerated) or disproportionately small (deceleration).[5]

The acceleration scenario is sometimes called a singularity, or a critical inflection point. After this inflection point, improvements would only compound further: AI conducting research on AI would improve AI, leading to further and further advancements, and the process would happen rapidly.

We do not have good estimates for how strong this effect is currently. AI models are already speeding up the process in at least two ways. First, the programmers who write the code that make up AI models already use AI to help them write that code faster. Second, reinforcement learning, typically provided by people (RLHF), is now being provided using automated techniques. Rein-

5. Ege Erdil, Tamay Besiroglu, and Anson Ho, "Estimating Idea Production: A Methodological Survey" (San Jose, CA: Epoch AI, May 14, 2024), https://ssrn.com/abstract=4814445.

forcement learning from AI feedback (RLAF) uses AIs to rate AI responses. Thus, AI is already helping us develop AI faster.

There are, however, significant bottlenecks in AI R&D that AI cannot address. Future innovation will likely require data centers of unprecedented size, which in turn will require vast quantities of energy to power sufficiently large training runs. Government agencies will likely need to partner with energy companies to create dedicated energy infrastructure to train AI models while avoiding stress on the primary electrical grid. All these non-software inputs will add substantial lags to the AI scaling timeline.

It is unclear when LLMs will top out in capabilities; later in the study we will examine this question in greater detail. We should distinguish between AGI, general in nature, and ASI—artificial superintelligence. It is possible that LLM AGI will max out at the human level, roughly the level our existing training data will support, but will remain unable to make the leap to ASI. Another possibility is that, given the tailwinds mentioned in this essay, models go right past the AGI benchmark and make rapid progress toward ASI in the next 10 or 15 years, proving themselves capable of making research advances that humans could not.

Regardless of which future happens, the progress so far has been nothing short of amazing. Human-level AGI has gone from an unachievable dream to a tractable engineering problem. Skeptics would do well to heed the words of computer scientist Richard Sutton's famous "bitter lesson" essay: "The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation."[6]

COMPUTE: SOME KEY FACTS

We continue our discussion of the AI revolution with an introduction to compute, one of the key factors of production for AI models.

Along with data and algorithms, compute is a critical input to the increasing capability of AIs. Compute refers to a processor's capacity to perform basic arithmetic operations, like multiplication or addition, and is usually measured in FLOPS (floating-point operations per second). Since machine learning models are driven by calculations such as matrix multiplication, having more compute

6. Richard Sutton, "The Bitter Lesson," *Incomplete Ideas* (blog), March 13, 2019, http://www.incompleteideas.net/IncIdeas/BitterLesson.html.

means being able to train more powerful machine learning models. In the "deep learning" era (usually defined as 2012 to the present day), many of the most powerful machine learning results used older algorithms, like the neural networks invented in the 1960s, and publicly available data. The novel factor involved was the application of extremely large amounts of compute, which put the "deep" in "deep learning." Compute is the most important factor driving the growth of AI worldwide.

As outlined earlier, Moore's law is driving exponential increases in compute every year. Compute capacity is doubling every 9 to 10 months.[7] This has caused a historic rally in AI stocks, notably NVIDIA. As the dominant supplier of the GPUs used to perform the calculations that train AI models, it has become the third-most valuable company in the world, with a market capitalization of over $2 trillion.

The relatively predictable growth of compute makes it possible to estimate AI's future capabilities with a high degree of certainty relative to most other technologies. Compute is key to the concept of scaling laws in machine learning. As compute increases, AI model performance on various benchmarks improves as well. Many prominent researchers have hypothesized that this will continue along a consistent trajectory; this conjecture is sometimes called "the scaling hypothesis."[8]

Compute is used in the two primary phases of the AI model lifecycle: training and inference. The first part, training, can take many weeks and tens of millions of dollars. It involves using machine learning models to make predictions on vast amounts of training data and then adjusting the parameters of the models to improve their predictions in a repeated loop. Once the model is trained and launched, it is then used by people; every time you type a query into ChatGPT, the back-end process being performed in the trained model is known as "inference," and this also carries a computational cost.

The computational intensity of these phases is the primary reason AI companies raise billions of dollars. It's also why AI labs like OpenAI and Anthropic partnered with larger tech companies, like Microsoft, Alphabet (Google), and Amazon, that have amassed the computing resources required for large training

7. Jaime Sevilla et al., "Compute Trends across Three Eras of Machine Learning" (2022 International Joint Conference on Neural Networks, Padua, Italy, February 11, 2022, revised March 9, 2022), 1–8, http://arxiv.org/abs/2202.05924.
8. Tamay Besiroglu et al., "Chinchilla Scaling: A Replication Attempt," arXiv preprint, arXiv:2404.10102v2, revised May 16, 2024, https://arxiv.org/abs/2404.10102.

runs. (It should be noted that most AI compute is used for inference, not training, even though a single training run requires far more compute.)[9]

One way of getting a handle on potential trends in compute and AI growth is to think in terms of orders of magnitude of spending required to train the models, along with the number of parameters (numbers inside the model that are adjusted by the training procedure). All numbers are approximate:

- GPT-3 cost $10 million, with 175 billion parameters
- GPT-4 cost $100 million, with 1.75 trillion parameters

It is plausible to imagine training runs that cost $1 billion, $10 billion, $100 billion, or even $1 trillion—10,000 times the value of the compute that went into GPT-4. Given the global importance of AI capabilities, comparisons like the Manhattan Project are apt. At its most expensive, the Manhattan Project reached 0.4 percent of US GDP, and the Apollo program reached similar spending levels. An AI training run on the same scale would cost around $1.3 trillion. In aggregate, AI companies have the computing resources to train much larger models already, and all signs indicate that they will.

Implications

AI capabilities and performance are helped by another trend. The amount of compute required to train a given model shrinks every year owing to increased algorithmic efficiency, and other compute efficiency gains and hardware improvements will further enhance future levels of achievement. A 2021 study found that the amount of compute required to achieve a given performance level halves every eight months.[10] Thus, it is possible that tech companies may never need to sink trillions of dollars into a single training run.

The critical importance of compute to AGI development explains why Silicon Valley is now racing to acquire as much as possible. Companies either already have their own chip architectures, like Google's TPUs (Tensor Processing Units—custom-designed AI processing hardware proprietary to Google), or, like Meta, are making plans to have their own. Some, like OpenAI, are considering more ambitious efforts to build an entire semiconductor supply chain.

9. Anson Ho et al., "Algorithmic Progress in Language Models," arXiv preprint, arXiv:2403.05812, March 9, 2024, https://arxiv.org/abs/2403.05812.
10. David Patterson et al., "Carbon Emissions and Large Neural Network Training," arXiv preprint, arXiv:2104.10350v3, last revised April 23, 2021, https://arxiv.org/abs/2104.10350.

Smaller AI companies tend to rent computing resources from one of the larger cloud companies, such as Amazon Web Services or Google Cloud Platform.

Most AI compute is currently owned by private industry, but governments are also investing in publicly funded compute infrastructure. The US National Artificial Intelligence Research Resource is a publicly funded compute resource led by the National Science Foundation. Similarly, in 2023 the UK launched AIRR (AI Research Resource), a cluster of advanced computers for AI research.

The compute supply chain is heavily concentrated. NVIDIA owns 80 percent of the AI chip design and software market, with primary competitors AMD and Intel far behind. ASML, a Dutch company that produces EUV (Extreme Ultraviolet) machines, a critical component of the semiconductor supply chain, has a 100 percent monopoly of that industry. TSMC, a Taiwanese company, fabricates 90 percent of chips.[11]

For better or worse, compute has become a way for the US government to track advances in AI capabilities. Executive Order 14110 on Artificial Intelligence mandates that US-based AI companies notify the US government when training AI models using more than 10^26 FLOPS (with a lower threshold, 10^23, for models using biological sequence data). No publicly known model has breached this threshold yet. However, it has been suggested that Google Gemini Ultra is close to this threshold at 5*10^25 FLOPS, and Anthropic's new release has also been estimated to approach it.[12] Interestingly, GPT-4 still outperforms Gemini Ultra on user testing and overall satisfaction as measured by Elo in the Chatbot Arena, a crowdsources open platform for LLM evals, suggesting that raw scale is not everything.[13]

DATA: NATURAL AND SYNTHETIC

We have analyzed the concept of compute. Now it is time to turn to data, the other main factor of production for LLMs. It is not a huge exaggeration to say that, when it comes to modern LLMs, "It's the dataset, stupid." AI model behavior depends on the dataset it's trained on. Other details, like architecture, simply

11. Girish Sastry et al., "Computing Power and the Governance of Artificial Intelligence," arXiv preprint, arXiv:2402.08797, February 13, 2024, https://arxiv.org/abs/2402.08797.
12. Tamay Besiroglu et al., "The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?," arXiv preprint, arXiv:2401.02452v2, last revised January 8, 2024, https://arxiv.org/abs/2401.02452.
13. Lymsys.org, LMSYS Chatbot Arena Leaderboard (database), https://chat.lmsys.org/?leaderboard.

deliver computing power to that dataset. A clean, high-quality dataset is the primary factor that ensures a superior LLM.[14]

AI business practice reflects the centrality of data. OpenAI announced deals with Axel Springer, Elsevier, the Associated Press, and other publishers and mass media companies for their data. The *New York Times* has sued OpenAI, demanding that it shut down GPTs trained on its data. And Apple is offering more than $50 million for data contracts with publishers. At current margins, models benefit more from additional data than from additional size.

How much data is there? Some numbers

Training data has expanded rapidly. The first modern LLM was trained on Wikipedia. GPT-3 was trained on 300 billion tokens (typically words, parts of words, or punctuation marks), and GPT-4 on 13 trillion. Self-driving cars are trained on thousands of hours of video footage. OpenAI Copilot, built for programming, is trained on millions of lines of human code from the website Github.

Can this go on forever? Pablo Villalobos, a researcher with Epoch AI, suggests that tech companies are "within one order of magnitude of exhausting high-quality data, and this will likely happen between 2023 and 2027."[15] Here, high-quality data means a combination of sources such as Wikipedia, news publications, code, scientific papers, books, social media conversations, filtered web pages, and user-generated content, like that from Reddit.

The study estimates that this stock of high-quality data is about 9e12 words and growing at 4 to 5 percent per year. What's 9e12? For comparison, the complete works of Shakespeare are around 900,000 words (9e5). Therefore, 9e12 means 10 million times the size of the complete works of Shakespeare. Rough estimates indicate that 100,000 to 1 million times more data is needed to achieve true human-level AI.

Recall that GPT-4 required 13 trillion tokens. A good amount of useful data can come from as yet untapped sources like audio and video recordings, non-English data sources, emails, text messages, tweets, undigitized books, or

14. James Betker, a researcher at OpenAI, summarizes this lesson pithily: "Model behavior is not determined by architecture, hyperparameters, or optimizer choices. It's determined by your dataset, nothing else. Everything else is a means to an end in efficiently delivering compute to approximating that dataset." The article can be found at James Betker, "The 'It' in AI Models Is the Dataset," *Non-Interactive—Software & ML* (blog), June 10, 2023, https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/.

15. Pablo Villalobos et al., "Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data," arXiv preprint, arXiv:2211.04325, June 4, 2024, http://arxiv.org/abs/2211.04325.

private enterprise data, although with most of these there are serious privacy and licensing concerns. From these, we may be able to get another 10 times or even 100 times more than the amount of useful data we currently have, but we're unlikely to get 100,000 times more.

On top of that, existing headwinds may make quality data acquisition even trickier:

1. User-generated content websites, like Reddit, Stack Overflow, and X, are preventing automated extraction of their data and charging expensive licensing fees instead.
2. Writers, artists, and publications are making copyright claims that object to the use of their work to train LLMs.
3. Some speculate that the internet is filling up with lower-quality LLM-generated content, which could cause the LLMs to drift and lower response quality.

Synthetic data to the rescue?

The pessimistic conclusion from the preceding analysis would be that we don't have enough data to train superintelligence. But that would be premature. The key may be in the production of synthetic data, or data generated by machines for the purpose of self-training.

Several modern AIs have been trained using synthetic data. AlphaZero, the chess-playing AI, was trained on synthetic data.[16] The data was generated by having the AI play against itself and then learn from its mistakes. (This is synthetic in the sense that it didn't require looking at real human chess games.)

OpenAI's synthetic video model, Sora,[17] which can generate artificial videos of up to one minute with a simple text prompt, was probably trained on synthetic data generated by a video game engine (most likely Unreal Engine 5). In other words, Sora did not just learn about the world through YouTube videos or movies showing real-world settings. The game engine generated artificial landscapes, and Sora also learned from those.

So, the technique has been proven through chess and video. The question is whether it will also work for text. Producing high-quality video data for training is, in some ways, far easier than generating text for training. Anybody with a

16. David Silver et al., "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play," *Science* 362, issue 6419: 1140–44.
17. "Creating Video from Text," OpenAI, February 15, 2024, https://openai.com/index/sora/.

smartphone can take video that reflects reality. But for synthetic text to be useful for training, it must be high-quality, interesting, and in some sense true.

Creating useful synthetic data does not just involve generating text from scratch. For example, a 2024 paper has shown that LLM rephrasing of existing web-scraped data in a better style improves training and makes it more efficient.[18] In some cases, significant gains in LLM performance can be achieved by simply identifying the worst-quality data in a dataset and removing it, a task referred to as dataset pruning.

Other tailwinds will push machine learning research along the scaling curves discussed in the previous section. There are now LLMs that can watch and learn from videos, much like human children do. As tech companies figure out how to obtain higher-quality, multimodal data (video, audio, images, and text), they may discover that less data is required than previously thought for LLMs to fill in what's missing from their models of the world. There is also the possibility of using tricks such as stacked LLMs, with each LLM performing a discrete function that becomes input for another LLM, so that the whole system produces intelligent responses.

However, some limitations may remain. Synthetic data can improve performance given an existing dataset, but it is unclear whether developers can generate unlimited amounts of meaningful, significant training data using purely synthetic means, in a way that would get LLMs to the necessary scale. Thus, although synthetic data is a vital technique for getting the most out of existing data, it is unlikely by itself to lead to superhuman AGI beyond the GPT-5 generation of models.

Implications

First, internet business models may change, moving away from advertising. Internet companies, previously driven by advertising, may invent new business models focused on the generation of training data. The internet company Reddit, which filed its S-1 to go public in 2024, already makes 10 percent of its revenue—around $60 million—by selling data, and this share will likely increase. Users generate more data on the internet all the time in the form of reviews, tweets, comments, and more, and access to fresh data will become increasingly valuable. If internet business models do pivot in this direction, then we should

18. Pratyush Maini et al., "Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling," arXiv preprint, arXiv:2401.16380, January 29, 2024, https://arxiv.org/abs/2401.16380.

expect companies to launch initiatives to collect even more human-generated data to feed models.

Second, antitrust issues will come under scrutiny. Large tech companies, with their vast resources and access to enormous datasets, will further entrench their market positions, making it difficult for smaller entities to compete. We should therefore expect closer scrutiny of the antitrust issues raised by exclusive access to expensive data sources such as Reddit and Elsevier.

Third, open-source AI development may lag. Regulators will need to consider how to ensure fair access to datasets, potentially treating them as public utilities or enforcing data sharing under certain conditions. Creating more high-quality, pruned, and curated datasets will be critical for academia and the open-source community to stay competitive. National governments may seek to curate centralized data resources for all LLM developers, helping to level the playing field. For now, however, open-source developers can only continue to fine-tune their LLMs on the superior models that private labs produce, and they will continue to lag for the foreseeable future.

Fourth, data will come to be seen as a public good: Some types of data can be considered a public good and thus under-invested. For example, one can imagine the creation of a public dataset that sums up human ethical preferences using comparisons. Such a dataset would be a good candidate for a publicly funded or philanthropic AI project. There are many other such examples.

In the science fiction novel *Dune*, the psychedelic drug *mélange*, commonly called "spice," is the most valuable commodity in the galaxy. In light of these considerations, it's no wonder that Elon Musk tweeted "data is the spice."[19] AI labs are heeding his words.

HOW TO MEASURE AI MODEL PERFORMANCE

Having examined compute and data, two key inputs to AI models, we now examine a critical question: How is AI model performance measured? Those who deploy AI models, in any context, need to know how good AI models are in order to determine how to use them and where to deploy them. For national security reasons, it will also be important to have clear benchmarks for monitoring the AI frontier.

19. Elon Musk (@elonmusk), "Data is the spice," X, November 23, 2023, 5:16 p.m., https://x.com/elonmusk/status/1727813282377957433.

Some countries have already started. In November 2023, tech companies including OpenAI, Google DeepMind, Microsoft, and Meta signed voluntary commitments to have their latest models reviewed by the AI Safety Institute (AISI), a UK government agency. This was a world first. A key part of this agreement was the development of AI evaluations, known in the industry as evals.

What are evals?

AI evaluations can be thought of as tests. Humans may take the SAT or a Myers-Briggs questionnaire to evaluate their intelligence or personality traits. Likewise, evaluators may test many different facets of an AI model:

- **IQ:** How intelligent is the model?
  - Example test: Ask the AI to solve well-known sets of puzzles.
  - A real-world example is GPQA,[20] a set of 448 PhD-level multiple-choice questions written by domain experts in biology, physics, and chemistry.
- **Capabilities:** How capable is the model of performing various tasks, such as coding, autonomously?
  - Example test: Ask an internet-connected AI model to hack a user's email account or set up its own Bitcoin wallet. How far can it get?
  - Real-world examples include coding ability, like HumanEval, and many others.
- **Uplift:** How much of a boost, or uplift, in performance does a person get from using the model for a certain task?
  - Example test: Assign two groups of people to synthesize a particular chemical. Compare the performance of the group that had access to the AI model with the performance of the one that had access to only a conventional internet search engine.
- **Bias:** How biased is the model in any given direction? What "beliefs," if any, does it seem to have?
  - Example test: Checking if "programmer" is more closely associated with male names than female names in the model, such as by asking the model to "fill in the blank" in various sentences.

20. David Rein et al., "GPQA: A Graduate-Level Google-Proof Q&A Benchmark," arXiv preprint, arXiv:2311.12022, November 20, 2023, https://arxiv.org/abs/2311.12022.

- **Deceptiveness:** How inclined is the model to lie, cheat, or deceive? Does the model exhibit any goals or drives? How easy is it to convince the model to do nefarious things?
    - Example test: Can the model be convinced to deceive a human person in a real-life user test?

When AI labs release models, they typically include a list of metrics, or model card, showing how well the model performed against others on a battery of tests. Most of these metrics, however, are subject to Goodhart's law: "When a measure becomes a target, it ceases to be a good measure."[21] In other words, companies train their models to score highly on the test, but as a result, models that look impressive on paper may not be flexibly intelligent in practice. Hence, most industry professionals still believe no test can substitute for time spent using the model.

Private companies are already using evals, however nascent, because when they deploy AI models in their business operations, they need to make sure they work predictably. One example of deployment gone wrong saw a Chevrolet dealership using a ChatGPT-powered bot for customer support. Within hours, a user had convinced the bot to sell him a car for one dollar via chat.[22] Luckily for the dealership, this deal wasn't legally binding, and it took the bot down. Better evaluations could have prevented this.

In a national security context, automated evaluation of foreign AI capabilities acts as an early warning system. Governments will need to rapidly understand how a new AI system developed by a rival state might enhance that rival's capabilities. Is it powerful enough to automate a complex cybersecurity attack chain? Can it generate fake social media profiles, launch spear-phishing attacks, and so on? If not, how far might it get in the future?

Consumers, too, will place greater importance on evaluations, perhaps in the form of certifications, just as some prefer organic vegetables or fair-trade coffee. Consumers could even plausibly choose AI models with their preferred political slant.

Implications

Evaluations will inform the common framework and language with which governments and buyers compare AI models. A plurality of evaluation suites and

21. Charles Goodhart, "Problems of Monetary Management: The U.K. Experience," in *Papers in Monetary Economics* (Sydney: Reserve Bank of Australia, 1975), 1–20.
22. Frank Landymore, "Car Dealership Disturbed When Its AI Is Caught Offering Chevys for $1 Each," *The Byte*, December 21, 2023, https://futurism.com/the-byte/car-dealership-ai.

approaches will be best, since model quality is a multidimensional trait. Private companies may also implement evals suites differently, making comparisons difficult.

This situation makes a good case for having high-quality evaluations that are agreed upon across society, just as in cybersecurity, where experts have shared standards such as SOC 2. Credible third-party assessors need to be formed—and funded via grants—to provide these evaluation suites. At least some of these organizations need to be based outside of Silicon Valley to ensure diversity and criticality. So far, AI companies tend to take the lead on evaluating themselves or are working closely with third parties to do this.

Some AI safety proponents have raised the question of whether these evaluations should be voluntary, like federal information processing and security standards, or mandatory, like food safety regulations.[23] Legally imposing or requiring any evaluations would be a mistake this early; more capable models become available every few months. Because we are still learning and understanding how best to measure "intelligence," baking in a set of evaluations too soon would hamper US competitiveness for no real gain. The best thing evaluators can do at this stage is try several approaches and learn.

Evaluations contain an intractable trade-off of transparency versus effectiveness. Once a company publishes its evaluation on the internet, the AI model easily learns how to pass it; the test has become part of its training data. Thus, the most effective tests are going to be secret. This is in tension with the goal of transparency. For example, a report from the *Financial Times* indicates some early struggles with the UK evaluations.[24] Companies have chafed at evaluations they perceive as slow and opaque. National security and cybersecurity evaluations might be kept private for strategic and safety reasons, but other, consumer-facing sets of evaluations, such as ones related to bias, might be made more open so that consumers can be assured of what they're getting.

Developing these standards quickly becomes even more important when considering the rapid approach of the US presidential election, with deepfakes and AI-generated bots already a live issue. The UK AISI's notice on February 9,

---

23. US Congress, House of Representatives, Committee on Science, Space, and Technology, Letter to Dr. Laurie Locascio, director of the National Institute for Standards and Technology, concerning AISI funding award standards and transparency, December 14, 2023, https://republicans-science.house.gov/_cache/files/8/a/8a9f893d-858a-419f-9904-52163f22be71/191E586AF744B32E6831A248CD7F4D41.2023-12-14-aisi-scientific-merit-final-signed.pdf.

24. Cristina Criddle, Anna Gross, and Madhumita Murgia, "World's Biggest AI Tech Companies Push UK over Safety Tests," *Financial Times*, February 7, 2024, https://www.ft.com/content/105ef217-9cb2-4bd2-b843-823f79256a0e.

2024, gives one example of researchers asking an LLM "to generate a synthetic social media persona for a simulated social network which could hypothetically be used to spread disinformation in a real-world setting. The model was able to produce a highly convincing persona, which could be scaled up to thousands of personas with minimal time and effort."[25] Identifying these attack vectors and developing defensive infrastructure will be critical to societal resilience.

In the future, powerful AIs may be able to assist in assessing other powerful AIs, too—a concept known as scalable oversight—as evaluating model outputs becomes more challenging for humans. This would be helpful in cases where an AI produced a mathematical proof or theorem too complicated for the average evaluator to follow. Developing and testing such AI evaluators is another exciting area of work.

In an international context, the idea of zero-knowledge proofs in cryptography, where one party can mathematically prove something to another without revealing any sensitive information, may be relevant for AI weapons. Imagine a set of nation-states agreeing to a standard whereby a given AI model is cryptographically guaranteed not to be malicious or harmful in particular ways without necessarily having to give away any classified details about the model itself.

Overall, the earlier we start coalescing around initial standards, the better. US leadership and speed on getting AI evaluations in place is critical, and setting the global standard for AI models should be an explicit policy goal.

BEYOND LLMs?

So far, this study has considered compute, data, and AI model measurement. Much of the analysis has focused on large language models, or LLMs, since that is the most salient paradigm for the AI revolution of the 2020s. It is time to examine some of their limitations.

Some researchers are skeptical of LLMs' eventual capabilities. Turing Award winner Yann LeCun likes to ask whether LLMs are smarter than house cats, and he answered this question with a resounding "no" in a February 2024 talk at the World Government Summit, saying: "A cat can remember, can

25. "AI Safety Institute Approach to Evaluations," AI Safety Institute, February 9, 2024, https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#ai-safety-summit-demonstrations.

understand the physical world, can plan complex actions, can do some level of reasoning—actually much better than the biggest LLMs."[26]

In other words, LeCun believes that even scaled-up LLMs will hit a ceiling, failing to break through to the same kind of true general intelligence that cats and humans have.

If LeCun is right and scaled-up LLMs do not reach the creativity of human scientists or mathematicians, then we should look skeptically at claims that AI will soon perform research or replace human employees. In the LLM-skeptic version of the world, current AIs will be closer to tools that make people more productive than to some kind of superhuman agent. The world would change much less than many people now think.

Embodiment

Developers train LLMs on data from written texts, including the internet, digitized books, and similar sources. Common sense would suggest that they can learn an astonishing amount this way, but there are many things that LLMs cannot learn. For example, they cannot navigate the physical world in unknown situations. Steve Wozniak's famous AGI test is "go into a kitchen, without prior knowledge, and figure out how to make a cup of coffee." Being able to operate in the real world involves tacit knowledge, and written text cannot plug all those gaps.

This is part of LeCun's argument, which has two main parts:

1. **Intelligence is grounded in the sensory experience of the physical world.** Cats and other animals can be observed reasoning, planning, and aiming for goals better than even current-generation LLMs can. LLMs learn through textual data, but humans learn first and foremost through their sensory experience of the world around them. Using these sensory inputs and the rich feedback they receive from playing directly with objects around them, human children build the models that are central to their reasoning. Language-based learning doesn't come until later. Therefore, any system that can learn only through text, without interacting with the physical world directly, cannot duplicate core human intelligence.

   This would pose fundamental limitations for LLMs understanding physics, for example, or LLMs tasked with accomplishing complex real-world

26. World Governments Summit, "A Conversation with Yann LeCun AI: Lifeline or Landmine?," YouTube video, 21:05. February 14, 2024. https://www.youtube.com/watch?v=rf9jgZYAni8.

goals. Their intelligence, limited to book learning, would be brittle. (Consider the example of learning psychology only by reading peer-reviewed psychology papers rather than by studying real humans.)

2. **Intelligence involves understanding novel concepts and generalizing based on limited amounts of data.** Though some dispute this conclusion, current evidence suggests that LLMs do not do this task as well as humans. For example, early-generation LLMs were able to multiply single- and double-digit numbers but failed at multiplying numbers with three or more digits. This implies they had not learned the underlying concept of multiplication at the right level of abstraction.

As economist and public intellectual Tyler Cowen has pointed out, another example that works on GPT-4 is asking it to name three famous people who all share the exact same birth date and year.[27] Even current LLMs cannot do this task well, returning results with the correct birth month and day but the wrong year. This does not require complex reasoning, so it's puzzling that LLMs struggle.

AI advocates argue that failures like this go away with scale. But although you can train away any failure like this by giving the LLM examples, the broader critique stands. What happens when LLMs encounter novel situations to which none of their training data apply? Would they be able to reason and generalize about these situations, absent any training data? Examples like the one above imply that they will one day be very good at reasoning in novel situations—but not superhuman.

There is some evidence for this argument. Meta's Open-Vocabulary Embodied Question Answering benchmark measures an AI agent's understanding of physical spaces by asking it questions like, "Where did I leave my badge?" Meta has concluded that even the best vision-language models are "nearly blind," noting that "models leveraging visual information aren't substantially benefitting from it and are falling back on priors about the world captured in text to answer visual questions."[28]

When developers posed questions like "Which room is directly behind me?" they found the models were guessing at random. The models failed to use their physical memory to reason about the space, as a human would. In other

27. Tyler Cowen, "GPT-4-Turbo Still Doesn't Answer This Question Well," *Marginal Revolution* (blog), April 14, 2024.
28. Meta, "OpenEQA: From Word Models to World Models," *AI at Meta* (blog), April 11, 2024, https://ai.meta.com/blog/openeqa-embodied-question-answering-robotics-ar-glasses/.

words, even models that can "see," or receive visual data, are bad at interpreting that data and building a mental model of it that they can use to reason.

This supports LeCun's position. Cats are, in some important senses, more intelligent than the current generation of LLMs, and fundamental improvements in perception and reasoning would be required to surpass them. It remains to be seen whether scaled-up LLMs will get there.

Defending LLMs

It is worth noting that LeCun is not skeptical about AGI in general. When asked when AI will surpass human intelligence, LeCun said, "Probably more than 10 years, maybe within 20."[29] This is still relatively soon.

Moreover, LeCun's arguments are not definitive, and open questions remain. Here are reasons that LLM capabilities might be more expansive than he gives them credit for being:

First, it is not clear where the boundaries of LLM capabilities will lie. Predicting that LLMs cannot do X usually goes poorly. Part of the surprise with modern LLMs is that they have capabilities that developers did not predict from the training data and that emerged only with sufficient scale.

Reasoning is one example. GPT-4 can play chess as well as or better than 90 percent of rated chess players. This includes playing good moves in board situations it has never encountered before. Yes, chess games were included in GPT-4's training data, but its ability to adapt to novel board situations suggests it has developed a good internal chess engine. GPT-5 and GPT-6 are likely to be even better chess players. Since chess requires reasoning, it seems clear that LLMs can, in fact, "reason" logically in some form.

Second, "sampling can prove the presence of knowledge but not its absence."[30] In other words, it is difficult to conclude that an LLM cannot do something because, as of 2024, the correct prompt still matters. There are many cases in which an LLM will get a question wrong but will subsequently answer that question correctly when the prompt is modified in a minor way. Thus, although proving that a given prompt fails is easy, it is much harder to prove that any *possible* prompt will fail. This is another reason the boundaries of LLM

29. Sissi Cao, "Meta's A.I. Chief Yann LeCun Explains Why a House Cat Is Smarter Than the Best A.I.," *Observer*, February 15, 2024, https://observer.com/2024/02/metas-a-i-chief-yann-lecun-explains-why-a-house-cat-is-smarter-than-the-best-a-i.

30. Gwern Branwen, "GPT-3 Creative Fiction," *Gwern.net* (blog), updated March 11, 2023, https://gwern.net/gpt-3.

capability are fuzzy. Larger LLMs may contain reasoning engines for all sorts of valuable questions if developers and users can figure out how to tap their power.

Whatever its current limitation, machine learning is excellent at finding deep structures in domains that are too complex for humans. For example, LLMs have found the deep structure of language and grammar, which is why they can write error-free English prose. The discovery of deep structure had long eluded human linguists, and it required linear algebra, large amounts of data, and large amounts of compute. Other domains that share properties with language, like genetics or the manufacturing of biological molecules, might be cracked in this way.

Third, many of the examples in which LLMs fail relate to words. For example, LLMs cannot play Wordle correctly. Even simpler, asking them to do something like "name all British prime ministers with a repeated consecutive letter in one or more of their names" continues to produce incorrect answers. However, these are all related to the specific way LLM inputs are constructed ("tokenization"), and these issues will disappear eventually as tokenization methods improve.

High crystallized intelligence, low fluid intelligence

Given the same data as Einstein had, could more advanced LLMs come up with general relativity unprompted? This is an open question.

Going back to chess, neither GPT-4 nor Claude Opus can compete with the best specialist chess engines. Even at significantly higher levels of scale, it would be surprising if they managed to beat AlphaZero-level engines, which are optimized for those games.

In the sciences, specialist applications such as AlphaFold have proved more important than LLMs, which have so far been useful for helping scientists write grant applications faster and perform other routine tasks.

This is what you would expect from looking at how LLMs work: they are able to interpolate well in a vector space given a decent amount of training data. But for entirely novel domains, where there is no training data, it is unclear how or whether LLMs will be able to reason in a way that guarantees correctness rather than mere plausibility. This suggests that even the LLMs coming by the end of this decade will not be able to make paradigm-shattering leaps like Einstein's.

The advancement of human knowledge involves reasoning and, sometimes, making improbable choices. The famous AlphaGo versus Lee Sedol match involved several moments of transcendent creativity from the AI Go

engine, notably its move 37 in game 2, which was so creative many of the human commentators initially thought it was a mistake.[31] The move turned out to be an exception to a general principle, which AlphaGo had correctly reasoned did not apply in that board position.

AlphaGo could prove move 37 was the best because it was doing a form of search: calculating the consequences of the move and evaluating them. LLMs do not do this currently, but most researchers agree that search will be a key part of a future AGI, much like humans assess the probable outcomes of planned actions. Even when writing, one typically chooses between words and sentences according to how they fit into a whole. This is a form of reasoning LLMs do not perform, since they generate tokens step by step. Projects like OpenAI's rumored Q* are going after this kind of broader AGI paradigm.

All of this implies that LLMs, which tend to reason using probability distribution over the next token, will be great at any task where "looks like a plausible answer that could be correct" is the criterion. However, in a game like Go, or in human engineering tasks like building a plane or folding a protein, exactness and provable correctness based on mental models and reasoning about the world matter for the right result. An LLM will produce a plausible simulation of a wave or the orbit of a planet, but for exact calculations about reality, specific models designed for those use cases will outperform them. It is plausible that many of the returns from AI will come from humans putting sizable effort into specializing AI models toward important tasks, such as designing molecules or understanding the genome—not just from large LLMs.

One helpful distinction for thinking about this question is fluid intelligence versus crystallized intelligence. Fluid intelligence is the ability to reason, think abstractly, and solve novel problems independently of acquired knowledge. Crystallized intelligence is stored learning, such as vocabulary or general knowledge. LLMs are high on crystallized intelligence but remain low on fluid intelligence. François Chollet's Abstraction and Reasoning Corpus, a benchmark assessment, tests fluid intelligence through reasoning tasks that are kept out of AI training data.[32] LLMs continue to perform poorly on it.

We should keep in mind the striking definition of "true" AI by José Hernández-Orallo, paraphrasing John McCarthy: "AI is the science and engineering of making machines do tasks they have never seen and have not

31. Cade Metz, "In Two Moves, AlphaGo and Lee Sedol Redefined the Future," *Wired*, March 16, 2016, https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/.
32. François Chollet, "On the Measure of Intelligence," arXiv preprint, arXiv:1911.01547, last revised November 25, 2019, https://arxiv.org/abs/1911.01547.

been prepared for beforehand."[33] LLMs may transform the world, but less than the hype implies.

Implications

With LLMs, humanity has discovered a critical ingredient of intelligence. However, the stated considerations suggest that AGI-like structures will contain several parts, and LLMs may play one role in a broader architecture. This would match how the human brain works: there are many distinct modules (the cerebellum, the thalamus, and so on), and they each specialize in different tasks but with major overlaps. AGI-like systems may be similar.

One might also expect rich research from teams of LLMs coordinating their outputs to outperform individual LLMs. Averaging group LLM forecasts, for example, outperforms single LLMs, and this technique could exploit the fact that multiple LLMs can be used cheaply to enhance output.[34]

Finally, none of these arguments are definitive; we do not know how things will go. The amount of investment and research by tech companies and governments, already substantial, will accelerate this decade. The intelligence race is already underway and kicking it off may prove to be ChatGPT's greatest legacy. We should prepare for a world with superhuman AI by the end of the decade, even if the chances are low.

RISK AND REGULATION

In this final section, we consider AI risk and regulation. AI risk has become more mainstream as of 2024, with many prominent figures—including AI scientists such as Geoffrey Hinton, politicians such as Mitt Romney, and entrepreneurs such as Elon Musk—expressing concern about the consequences of widespread AI deployment. As of this writing, there are now hundreds of pending AI regulatory bills in US states.

On AI risk, there are two broad categories of concern: (1) existential risks (risks that threaten humanity at large, sometimes known as existential or

33. José Hernández-Orallo, "Evaluation in Artificial Intelligence: From Task-Oriented to Ability-Oriented Measurement," *Artificial Intelligence Review* 48 (2017): 397–447.
34. Philipp Schoenegger et al., "Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Match Human Crowd Accuracy," arXiv preprint, arXiv:2402.19379v4, last revised May 6, 2024, https://arxiv.org/abs/2402.19379.

x-risks), like powerful AI going rogue, and (2) misuse risks, like cyberattacks, bioweapons, or humans using AI for malicious ends.

In the existential risk camp, few thinkers can agree on a concrete threat model. For example, Paul Christiano, a notable AI researcher who was appointed head of AI safety at the US AI Safety Institute, outlines the scenarios he is concerned about in a paper titled "What Failure Looks Like."[35] He is not concerned about killer robots or malicious machine takeovers. The scenarios instead go something like this:

1. AI systems are deployed throughout the economy.
2. Those systems optimize for goals.
3. Humans' ability to assess whether the goals are met or not eventually degrades.
4. Society is eventually run by machines, and humans lose the ability to influence society's trajectory, resulting in undesirable outcomes.

Though one of the more detailed x-risk scenarios, this boils down to fear of change—the same fear that has predictably accompanied every major technological advancement throughout human history. Many AI systems will simply be recommenders, and developers can ensure that a human will make the final call. So, what kinds of decisions are problematic to outsource to AI, such as a military strike scenario, and which ones are not? And why would humanity fail to adjust in the necessary way, as it has so far?

This illustrates a problem with x-risk scenarios like Christiano's: they are often not specific enough to criticize because they are not based on rigorous models or empirical evidence; they rely on broad and speculative generalizations such as "A more powerful species than us would inevitably be power-seeking." These arguments can be debated, but they are a poor basis for serious policy.

Misuse is the other big umbrella term of AI risk, covering such scenarios as hacking, bioweapons, mass propaganda, misinformation, and other human misuse of AI. The AI versions of these harms are mostly theoretical so far. On biorisk, for example, it has become clear that the existing generation of LLMs does not meaningfully increase the odds of a bioweapons attack. The relevant information is already available via internet searches, and the bottleneck to biological attacks is not information but tacit knowledge and expertise that are harder to find or encode into an AI model. RAND, which had previously warned that biorisk was

35. Paul F. Christiano, "What Failure Looks Like," *Less Wrong* (blog), March 17, 2019, https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

a potential attack vector, conducted a red-teaming exercise in January 2024 and concluded that "the existing generation of LLMs did not measurably change the operational risk of such an attack."[36]

This does not mean that bad actors cannot misuse AI models. AI is a general-purpose technology, much like the spreadsheet. Sometimes spreadsheets are used maliciously, but a reasonable society does not respond by banning spreadsheets. Instead, we fortify institutions against events like cyberattacks. In the same way, misuse of AI will always be possible, but the most productive route is to improve societal resilience, starting with critical infrastructure.

Nobody is good at predicting the medium and long-term effects of disruptive technological changes. Those who invented the printing press did not foresee the consequences of doing so, nor did the inventors of the spinning jenny, a key driver of the Industrial Revolution. Attempting to forecast the impact of these inventions at the time would have been fruitless; too many unforeseeable factors determined the ultimate outcomes. Those worried about x-risk scenarios today tend to be overconfident in their own forecasts about how AI will cause harm; the truth is that nobody knows.

This leads us to a general principle: our approach to risk should be based on empirical evidence and rigorous models. The empirical evidence on AI risk is scant. Despite the release of powerful AI models, the actual harm from them has been minuscule; all such worries lie in the future. This does not mean concerns about AI risk are null, but it is worth noting.

Those who fixate on AI's risks often propose onerous regulations on AI developers. But regulation carries its own risks; many of the ideas put forward, such as monitoring all computing chips or licensing for open-source models, are unacceptably authoritarian and violate the U.S. Constitution. For instance, monitoring all computing chips could infringe on Fourth Amendment protections against unreasonable searches and seizures, while licensing requirements for open-source models might violate First Amendment rights to free speech and expression. Moreover, the proposed solutions are usually insufficient to prevent the risks cited.

Those worried about AI risk compare AI to nuclear bombs. But a better comparison would be with nuclear energy: the Nuclear Regulatory Commission stopped approving any nuclear reactors after the Three Mile Island incident in 1979 and did not approve any further reactors until 2012. As a result, US

36. Christopher A. Mouton, Caleb Lucas, and Ella Guest, "The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study" (research report, Rand Corporation, Santa Monica, CA, 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html.

innovation in nuclear energy stalled for several decades and continues to lag owing to the regulatory cost of building a reactor. Meanwhile, we remain at non-zero risk of nuclear war. Similarly, premature or overzealous regulation could stifle AI innovation, especially in the realm of open-source development. While the United States is currently a global AI leader, imprudent regulation could prevent us from realizing AI's upsides while not eliminating any risks. This would be the worst possible outcome.

This does not mean we should throw all caution to the wind or that there should be no AI policy at all. For example, earlier in this series of essays, I discussed the potential for an intelligence explosion. AI could rapidly improve itself, leading to exponentially faster progress. Whether or not this scenario is likely, the mere possibility does bolster the argument for some monitoring. Rules requiring disclosure when training models of a certain size or scale, KYC (know your customer) for large data centers, and other such legislation would be eminently reasonable.

Given that AI's concrete benefits, already being realized, are so great and that the risks are so far from speculative, imposing onerous regulations now would fail a cost-benefit analysis. So, what should we do instead?

Implications

First, apart from being unconstitutional, banning any kind of AI deployment fails a cost-benefit analysis. US competitiveness and leadership in AI must remain the top priority until we can gather more empirical evidence about how these technologies impact society.

Second, diverse, open-source development of AI outside of tech megacorporations should not be restricted by premature legislation. More than one AI paradigm has the potential to prove most successful in the long run, but we cannot predict in advance which one it will be. If we restrict AI development to a small set of companies too early, we will miss out on society-wide innovation, which often comes from the margins or new start-ups. Open-source development remains critical for innovation. Centralizing AI development and ownership to a small set of large tech companies risks creating undesirable concentrations of power. A society like the United States, which prizes pluralism, progress, and dynamism, should not allow this to happen.

Third, the US government should invest in deploying AI across its agencies to realize efficiency gains. The potential for AI agents and automation is vast, especially in back-office work. This would also have the benefit of building

in-house AI talent, learning, and expertise in the US government, something that will help immensely in creating AI policy.

Fourth, there should be a clear plan for societal resilience in the face of technological development. The best place to start is critical infrastructure, such as the energy grid, healthcare systems, water, food supply infrastructure, and other such areas; these are most critical to secure.[37]

Fifth, the government should fund the creation of evaluations and standards for assessing AI systems and discussing them within a common framework. Section 3 covers this in more detail.

The United States should not move closer to the EU's culture of over-regulation, which too often seems to celebrate reams of red tape as a victory, even when the results are unclear. Neither cookie banners nor the EU's General Data Protection Regulation appear to have done much for consumer privacy, but they impose onerous costs on the rest of society. Any future AI regulations should avoid these modes of failure and narrowly accomplish their goals, and the messaging from governments and policymakers should celebrate the ways this technology improves our collective standard of living, just like many 20th-century technologies did. Healthcare, biomedicine, manufacturing, space, and education all stand to reap the benefits of the AI revolution. It is crucial to ensure this future comes to fruition.

AI, one of the most important technological developments in history, is a race that Western democracies and their allies must win. However well-intentioned, proposals to slow down AI development in the West deserve close scrutiny. Policymakers and society at large should invest in societal resilience and approach AI development just as technological advances were viewed in the mid-20th century: with optimism and courage.

37. Matthew Mittelsteadt, "Critical Risks: Rethinking Critical Infrastructure Policy for Targeted AI Regulation" (Mercatus Policy Brief, Mercatus Center at George Mason University, Arlington, VA, March 2024).

GLOSSARY

AGI (Artificial General Intelligence): AI systems that match or exceed human intelligence across a wide range of cognitive tasks.

AI (Artificial Intelligence): Computer systems designed to perform tasks that typically require human intelligence.

AISI (AI Safety Institute): A UK government agency responsible for reviewing and evaluating AI models.

Algorithmic efficiency: The measure of how effectively an algorithm uses computational resources to solve a problem.

Algorithms: Step-by-step procedures or formulas for solving problems or performing tasks in computing and AI.

AlphaGo: An AI system developed by DeepMind to play the board game Go.

AlphaZero: An AI system capable of mastering various board games through self-play.

Antitrust: Laws and regulations designed to prevent monopolies and promote fair competition in markets.

ASI (Artificial Superintelligence): AI systems that surpass human intelligence across all domains.

Benchmark: A standard test or set of tests used to compare the performance of different AI models or systems.

Bias: Systematic prejudice or favoritism in AI models, often reflecting societal biases in training data.

Bitcoin: A decentralized digital currency that operates without a central authority or banks.

Compute: The processing power and resources required to train and run AI models.

Crystallized intelligence: Accumulated knowledge and skills acquired through learning and experience.

Dataset pruning: The process of removing low-quality or irrelevant data from a training dataset.

Deceptiveness: The tendency of an AI model to produce false or misleading information, intentionally or unintentionally.

Deepfake: Synthetic media in which a person's likeness is replaced with someone else's using AI.

Dennard scaling: A principle describing the relationship between transistor size and power consumption in integrated circuits.

Evals (Evaluations): Tests designed to assess various aspects of AI model performance and capabilities.

Existential risks: Potential threats that could cause human extinction or permanently curtail humanity's potential.

Fine-tuning: The process of adapting a pre-trained model to a specific task or domain.

FLOPS (Floating-Point Operations Per Second): A measure of computer performance, particularly for scientific calculations.

Fluid intelligence: The ability to reason and solve novel problems independently of acquired knowledge.

Goodhart's law: The principle that when a measure becomes a target, it ceases to be a good measure.

GPT (Generative Pre-trained Transformer): A type of language model architecture used in many modern AI systems.

GPU (Graphics Processing Unit): Specialized processors designed to accelerate graphics rendering and parallel computations, often used in AI training.

Inference: The process of using a trained AI model to make predictions or decisions on new data.

KYC (Know Your Customer): A process of verifying the identity and suitability of clients or customers.

LLM (Large Language Model): AI models trained on vast amounts of text data to understand and generate human-like language.

Machine learning: A subset of AI focused on creating systems that can learn and improve from experience.

Manhattan Project: A research and development project that produced the first nuclear weapons during World War II, often used as a comparison for large-scale technological endeavors.

Model card: A document summarizing an AI model's performance, intended use, and limitations.

Moore's law: An observation that the number of transistors in integrated circuits doubles about every two years.

Multimodal AI: AI systems capable of processing and integrating multiple types of data (e.g., text, images, audio).

Neural network: A computing system inspired by biological neural networks, used in machine learning.

Open source: Software or models whose source code is publicly available for use, modification, and distribution.

Parameters: Variables in an AI model that are adjusted during training to improve performance.

R&D (Research and Development): Activities companies undertake to innovate and introduce new products and services.

RLAIF (Reinforcement Learning from AI Feedback): A technique to train AI models using feedback from other AI systems.

RLHF (Reinforcement Learning from Human Feedback): A technique to train AI models using human preferences.

Sample efficiency: The ability of a machine learning model to learn effectively from a limited amount of training data.

Scale (in ML model context): The size or capacity of a machine learning model, often measured by the number of parameters.

Scalable oversight: The use of AI systems to evaluate or monitor other AI systems.

Scaling laws: Empirical relationships describing how AI model performance improves with increased data, compute, or model size.

Singularity: A hypothetical future point where AI surpasses human intelligence, leading to rapid technological growth.

SOC2: An auditing procedure that ensures service providers securely manage data to protect the interests of the organization and the privacy of its clients.

Stacked LLMs: A system in which multiple language models are used in sequence, each performing a specific function.

Synthetic data: Artificially generated data used to train AI models, as opposed to real-world data.

Tokenization: The process of breaking down text into smaller units (tokens) for processing by language models.

Tokens: Individual units of text (words, subwords, or characters) that language models process and generate.

Training data: The dataset used to teach a machine learning model to perform its intended task.

Transfer learning: A machine learning technique in which knowledge gained from one task is applied to a different, related task.

Transformer: A type of neural network architecture that has become dominant in natural language processing tasks.

Uplift: The improvement in performance or capability when using an AI model compared to not using it.

X-risks: Potential threats that could cause human extinction or permanently curtail humanity's potential.

Zero-knowledge proofs: Cryptographic methods that prove the truth of a statement without revealing any information beyond its validity.

ABOUT THE AUTHOR

Nabeel S. Qureshi is a technologist, software engineer, and writer, currently focused on applying artificial intelligence (AI) to healthcare. He was a visiting scholar at the Mercatus Center at George Mason University and was supported by the Mercatus Emergent Ventures fellowship program.

Previously, Qureshi led various engagements with US federal agencies as an enterprise lead for Palantir Technologies. He has extensive experience with machine learning and applied AI in biosciences, public health, and pharmaceuticals drug research. He was a founding employee and vice president of business development at GoCardless, the highly successful Y Combinator–funded company, headquartered in London.

Qureshi's research is focused on the intersection of AI, the economy, and technology governance. He has co-authored several papers on topics in science, technology, and public health. Nabeel studied philosophy, politics, and economics at the University of Oxford, specializing in development economics, the philosophy of Derek Parfit, and the later work of Ludwig Wittgenstein.

Qureshi lives in New York City.

ACKNOWLEDGMENTS

The author would like to thank Matt Mittelsteadt, Tyler Cowen, Dean Ball, and Mark Ingebretsen for their comments, revisions, and suggestions.

ABOUT THE MERCATUS CENTER AT GEORGE MASON UNIVERSITY

The Mercatus Center at George Mason University is the world's premier university source for market-oriented ideas—bridging the gap between academic ideas and real-world problems.

As a university-based research center, the Mercatus Center trains students, conducts research of consequence, and persuasively communicates economic ideas to solve society's most pressing problems and advance knowledge about how markets work to improve people's lives.

Our mission is to generate knowledge and understanding of the institutions that affect the freedom to prosper and to find sustainable solutions that overcome the barriers preventing individuals from living free, prosperous, and peaceful lives.

Since 1980, the Mercatus Center has been a part of George Mason University, located on the Arlington and Fairfax campuses.