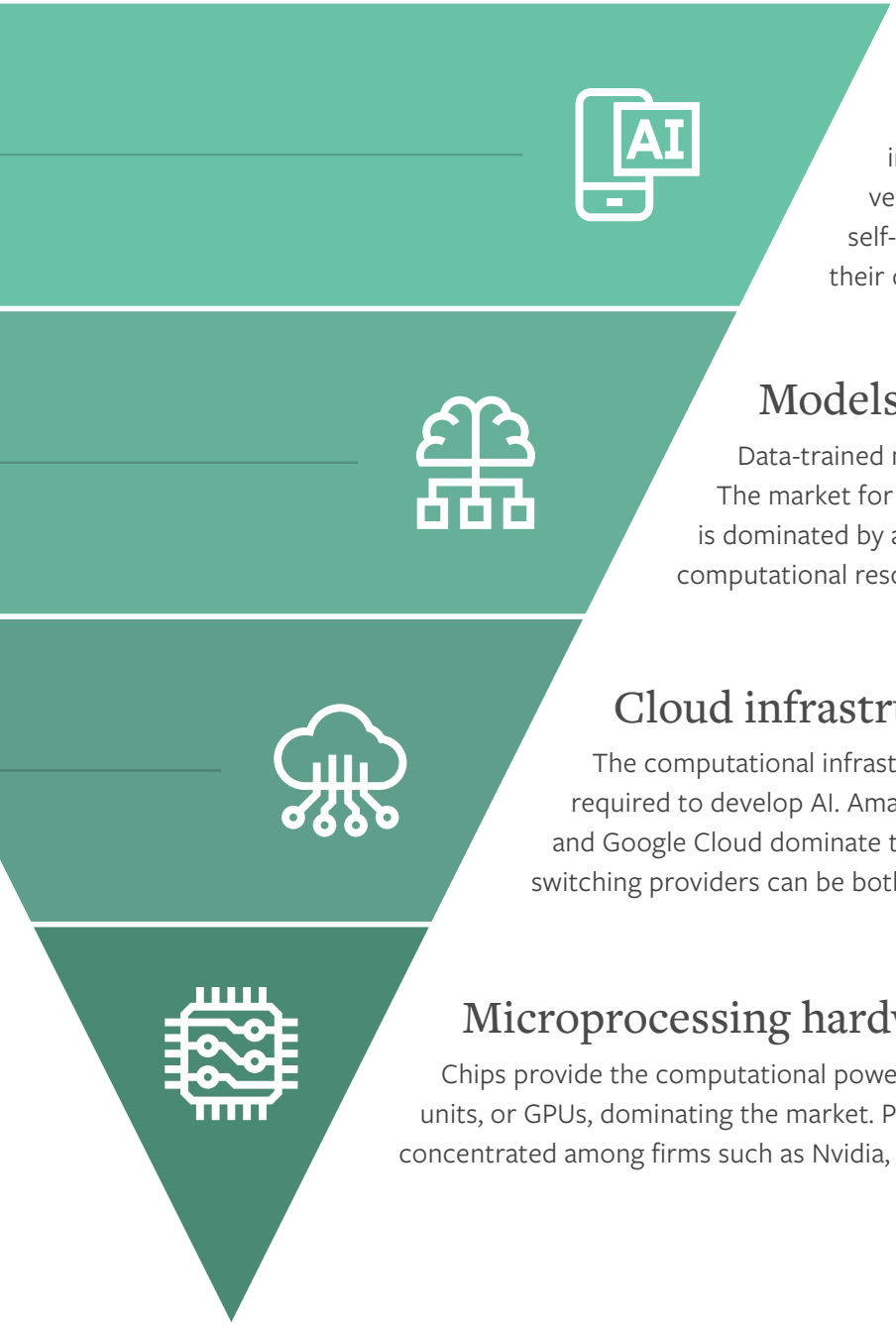
Washington Center for Equitable Growth
Evidence for a stronger economy

equitablegrowth.org

The AI Technology Stack

Artificial intelligence is more than just the applications we use, such as chatbots and predictive algorithms. Beneath the surface lies a technology stack that is increasingly concentrated at each layer. Regulation can address the concentration in this complex supply chain to promote innovation, competition, and fairness.

Tejas N. Narechania, University of California, Berkeley



Applications

Applications, such as ChatGPT, are how consumers interact with AI. Some applications benefit from vertically integrated ecosystems, giving rise to risks of self-preferencing by firms that develop models or build their own cloud-computing infrastructure.

Models

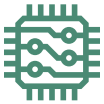
Data-trained models are the "brains" behind AI applications. The market for models tends toward concentration and thus is dominated by a few firms with control over training data and computational resources.

Cloud infrastructure

The computational infrastructure—servers and networks—required to develop AI. Amazon Web Services, Microsoft Azure, and Google Cloud dominate the market for cloud services, and switching providers can be both difficult and costly.

Microprocessing hardware

Chips provide the computational power for AI, with graphical processing units, or GPUs, dominating the market. Production in this layer is highly concentrated among firms such as Nvidia, TSMC, and ASML.



Microprocessing hardware

This is the base layer in the so-called AI technology stack, or the AI supply chain. This layer features microprocessors, primarily graphical processing units, or GPUs, which perform the computations necessary for AI. Key players, such as Nvidia, dominate chip design, while fabrication is concentrated in Taiwan Semiconductor Manufacturing Company, or TSMC. Specialized tools from firms such as ASML underscore concentration in the layer, which is partially driven by dependence on specific technologies. As noted, the hardware layer is highly concentrated, with Nvidia holding more than 80 percent of the AI market and TSMC leading in fabrication. High costs for equipment and production impose substantial barriers to entry. Policymakers should therefore consider subsidizing semiconductor production facilities to diversify supply chains and reduce reliance on a concentrated industry. Moreover, nondiscrimination rules could ensure fair access to hardware for downstream developers.

Cloud infrastructure

This layer involves the computational infrastructure—for example the networks, servers, and storage, as well as the datacenters that house them—for developing and processing AI data and models. Dominated by Amazon Web Services, Microsoft Azure, and Google Cloud, it provides the backbone for AI operations. The layer is oligopolistic, with high capital costs and switching barriers creating lock-in effects for consumers. Interoperability between providers is minimal. Policymakers should therefore consider building public cloud services to provide competitive alternatives and set fair pricing baselines. Clear interconnection and interoperability standards might help to reduce switching costs and foster greater competition, including by new entrants.

AI models

This layer encompasses the algorithms, data, and training processes that define AI's core operational capabilities. Models may be open-source or proprietary, and accessed via application programming interfaces, or APIs, or downloaded for local development from so-called model hubs. Here, vertical integration is key. Firms leverage their existing data assets and computational power to score advantages in the model layer. Smaller developers face barriers due to high costs and limited access to proprietary resources. Policymakers should therefore consider addressing vertical integration through, say, nondiscrimination rules that ensure equal terms for access to proprietary models, thereby limiting the power of dominant firms to favor their own applications.

AI applications

The application layer includes consumer-facing software, such as ChatGPT, which integrates outputs from the prior layers. While there is a thriving market for AI-driven applications, it is important to note that some applications benefit from vertically integrated ecosystems while others attempt to compete with vertically integrated applications. As noted above, policymakers should therefore take care to ensure that vertically integrated applications compete fairly with their counterparts and do not benefit from favorable terms for accessing models, data, and computational resources.