An Antimonopoly Approach to Governing Artificial Intelligence

Vanderbilt
Policy Accelerator
for Political Economy & Regulation

VANDERBILT UNIVERSITY

AN ANTIMONOPOLY APPROACH TO GOVERNING ARTIFICIAL INTELLIGENCE

Tejas Narechania & Ganesh Sitaraman[*]

Since OpenAI released ChatGPT in the fall of 2022, debates over regulating artificial intelligence (AI) among policymakers, technologists, and scholars have intensified. But for all the interest in regulating AI, there has been little discussion of AI's industrial organization and market structure. This is surprising because parts of the AI supply chain (i.e., the "layers" in the "AI technology stack") are already monopolistic or oligopolistic.

In this Article, we make the case for an antimonopoly approach to governing artificial intelligence. We show that AI's industrial organization, which is rooted in AI's technological structure, suffers from market concentration at and across a number of layers. And we argue that an AI oligopoly has undesirable economic, national security, social, and political consequences.

Our analysis of AI's industrial organization leads to some important conclusions: that important elements of AI are stable enough to invite regulation, notwithstanding ongoing technical development; that ex ante tools of competition regulation are likely to prove more effective than modes of ex post enforcement, as under antitrust law; that regulation can help facilitate more downstream innovation and that the current market structure may in fact inhibit innovation; and that some of the most prominent worries about AI—such as bias and privacy—might themselves be partly the result of market structure concerns.

In light of these conclusions, we show how antimonopoly market shaping tools—the law of networks, platforms, and utilities; industrial policy; public options; and cooperative governance—can all help facilitate competition and combat inequality. As policymakers debate governing AI early in its technological lifecycle, antimonopoly tools must be part of the conversation.

[*] Professor of Law, University of California, Berkeley, School of Law, and New York Alumni Chancellor's Chair in Law, Vanderbilt University Law School, respectively. Thanks to Sarah Myers West, Sabeel Rahman, Douglas Schmidt, Matt Stoller, and Jules White for helpful conversations and suggestions and to Mathew Cha, Ramsay Eyre, and Lynn He for research assistance.

AN ANTIMONOPOLY APPROACH TO GOVERNING ARTIFICIAL INTELLIGENCE

Tejas Narechania & Ganesh Sitaraman

CONTENTS

INTRODUCTION

Since OpenAI released ChatGPT in the fall of 2022, debates over regulating artificial intelligence (AI) among policymakers, technologists, and scholars have intensified. The Biden White House issued a "Blueprint for an AI Bill of Rights,"[1] and the European Parliament passed the A.I. Act to regulate risky uses of AI technology.[2] Sam Altman—OpenAI's Chief Executive—has endorsed greater regulation of AI systems, [3] while notable industry figures including Elon Musk, Steve Wozniak, and Gary Marcus, have gone so far as to call for a "pause" on AI development.[4] Scholars and commentators have discussed a wide range of problems with AI and proposed regulatory strategies to address those problems.[5] Notable books and articles cover algorithmic bias,[6] misinformation and disinformation,[7] algorithmic collusion,[8] labor displacement,[9] legal personhood for AI,[10] liability rules,[11] common law regulation,[12] explainability and transparency,[13] the

[1] THE WHITE HOUSE, BLUEPRINT FOR AN AI BILL OF RIGHTS (Oct. 2022), https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[2] Adam Satariano, *Europeans Take a Major Step Toward Regulating A.I.*, N.Y. TIMES (June 14, 2023), https://www.nytimes.com/2023/06/14/technology/europe-ai-regulation.html

[3] Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing* (May 16, 2023), https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html

[4] James Vincent, *Elon Musk and Top AI Researchers Call for Pause on 'Giant AI Experiments'*, THE VERGE (Mar. 29, 2023, 4:08AM CDT), https://www.theverge.com/2023/3/29/23661374/elon-musk-ai-researchers-pause-research-open-letter

[5] For an overview applying a range of existing legal principles to AI, see JACOB TURNER, ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE (2018).

[6] SARA WACHTER-BOETTCHER, TECHNICALLY WRONG: SEXIST APPS, BIASED ALGORITHMS, AND OTHER THREATS OF TOXIC TECH (2017); VIRGINIA EUBANKS, AUTOMATING INEQUALITY (2018), SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION (2018); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. 1, 6–11 (2018); Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).

[7] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016).

[8] Ariel Ezrachi & Maurice E. Stucke, *Artificial Intelligence & Collusion: When Computers Inhibit Competition*, 2017 U. ILL. L. REV. 1775.

[9] Daron Acemoglu & Pascual Restrepo, *Artificial Intelligence, Automation, and Work*, in THE ECONOMICS OF ARTIFICIAL INTELLIGENCE (Ajay Agarwal, Joshua Gans & Avi Goldfarb, eds) (2019).

[10] Lawrence Solum, *Legal Personhood for Artificial Intelligence* (1992)

[11] David C. Vladeck, *Machines without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117 (2014)

[12] Mariano-Florentino Cuellar, *A Common Law for the Age of Artificial Intelligence: Incremental Adjudication, Institutions, and Relational Non-Arbitrariness*, 119 COLUM. L. REV. 1773 (2019)

[13] Solon Barocas & Andrew Selbst, *The Intuitive Appeal of Explainable Algorithms*, 87 FORDHAM L. REV. 1085 (2018); Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017). For how algorithms intersect with governmental transparency, see Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1 (2019).

FTC's regulatory powers over AI systems,[14] the right to contest AI determinations,[15] AI and the administrative state,[16] AI and constitutional rights,[17] and AI's role in the use of international force,[18] among other concerns.[19]

For all the interest in regulating AI, there has been little discussion of AI's industrial organization and market structure.[20] This is surprising because parts of the AI supply chain (i.e., the "layers" in the "AI technology stack," to use the parlance of the sector) are monopolistic or oligopolistic.[21] Indeed, one scholar has described how machine learning—the algorithmic foundation for AI applications—has natural monopoly characteristics, even under narrow economistic definitions.[22] As in other areas, monopoly and oligopoly in AI can not only distort markets, chill investment, and hamper innovation, but also facilitate downstream harms to users and help accumulate private power in relatively few hands.[23]

In this Article, we make the case for an antimonopoly approach to governing artificial intelligence. We show that AI's industrial organization, which is rooted in AI's technological structure, suffers from market concentration at and

[14] Michael Spiro, *The FTC and AI Governance: A Regulatory Proposal*, 10 SEATTLE J. TECH., ENVTL. & INNOVATION L. 26 (2020)

[15] Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957 (2021)

[16] David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. ON REG. 800 (2020); Ryan Calo & Danielle Keats Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. 797 (2021)

[17] Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1875 (2020)

[18] Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 J. NAT'L SEC. L. & POL'Y 1 (2019)

[19] *See, e.g.*, Ryan Calo, *Artificial Intelligence Policy: A Primer and a Roadmap*, 51 U.C. DAVIS L. REV. 399 (2017); Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J. L. & TECH. 353 (2016)

[20] To the extent there has been, it has largely focused on semiconductor manufacturing, and to a lesser extent, cloud infrastructure provision. But even then, these concerns have not generally been considered in the context of AI specifically. Among the rare works to examine competition aspects of AI are C. Scott Hemphill, *Disruptive Incumbents: Platform Competition in an Age of Machine Learning*, 119 COLUM. L. REV. 1973, 1975–81 (2019); Amba Kak & Sarah Myers West, *AI Now 2023 Landscape: Confronting Tech Power*, AI NOW INSTITUTE (April 11, 2023), https://ainowinstitute.org/2023-landscape; Staff in the Bureau of Competition & Office of Technology, *Generative AI Raises Competition Concerns*, FED. TRADE COMM'N (June 29, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns. One notable work on the AI supply chain is Jennifer Cobbe, Michael Veale & Jatinder Singh, *Understanding Accountability in Algorithmic Supply Chains*, FAccT'23: Proc. 2023 ACM Conf. on Fairness, Accountability, and Transparency, June 2023, https://doi.org/10.1145/3593013.3594073.

[21] We recognize that the firms in the AI supply chain are in different sectors. Semiconductor firms, for example, need not produce their chips for AI. But AI depends on the inputs we describe, and as we show, many of these layers are vertically integrated, meaning that AI-based applications are dependent on the market structure and organization of these layers. *See infra* Part I, for further discussion.

[22] Tejas N. Narechania, *Machine Learning as a Natural Monopoly*, 107 IOWA L. REV. 1543 (2022).

[23] For a discussion of these problems in the e-commerce context, see Lina M. Khan, *Amazon's Antitrust Paradox*, 126 YALE L.J. 564 (2017).

across a number of layers. And we argue that an AI oligopoly has undesirable economic, national security, social, and political consequences. Our analysis of AI's industrial organization leads to some important conclusions: that important elements of the AI sector are stable enough to invite regulation, notwithstanding ongoing technical development; that ex ante tools of competition regulation are likely to prove more effective than modes of ex post enforcement, as under antitrust law; that regulation can help facilitate more downstream innovation and that the current market structure may in fact inhibit innovation; and that some of the most prominent worries about AI—such as bias and privacy—might themselves be partly the result of market structure concerns. In light of these conclusions, we show how antimonopoly market shaping tools—networks, platforms, and utilities (NPU) law, industrial policy, public options, and cooperative governance—can apply to aspects of the AI sector.

The starting point for our analysis is a detailed understanding of the AI technology stack, which, so far as we are aware, legal scholars have not outlined in detail. Drawing, in Part I, on accounts from industry investors and analysts, we describe AI's technology stack in four basic layers: microprocessing hardware, cloud computing, algorithmic models, and applications.

The hardware layer includes the production of microchips and processors—the horsepower behind AI's computations. This layer is extremely concentrated, with a few firms dominating important aspects of production. The cloud computing layer consists of the computational infrastructure—the computers, servers, and network connectivity—that is required to host the data, models, and applications that comprise AI's algorithmic outputs. This layer, too, is highly concentrated, with three firms (Amazon Web Services (AWS), Google Cloud Platform, and Microsoft Azure) dominating the marketplace.

The model layer is more complicated than the first two, as it includes three sublayers (and even more within those): data, models, and model access. One primary input for an AI model is data, and so the model's layer's first sublayer is data. Here, companies collect and clean data and store it in so-called "data lakes" (relatively unstructured data sources) or "data warehouses" (featuring relatively more structure). Foundation models (which are distinct from *all* models in general) comprise the second sublayer.[24] Models are what many think of as "AI." These models are the output of an algorithmic approach to analyzing and "learning"[25] from the inputs that begin in the data sublayer. This "training"[26] process is expensive, and so models can be intensely costly to develop. Lastly, the third sublayer consists of modes of accessing these models—model hubs and APIs (short for "application programing interfaces"). While model hubs and APIs both offer access to a foundation model, they operate quite differently from each other. Model hubs are

[24] Rishi Bomnasani et al., *On the Opportunities and Risks of Foundation Models*, Center for Research on Foundation Models, Stanford Institute for Human-Centered Artificial Intelligence (HAI), at https://arxiv.org/abs/2108.07258 (manuscript)

[25] Narechania, *supra* note xx, at 1550 n.25, 1551 n.35 (2022) (on the use of terms such as "training" "learning" and "understanding" in the context of machine learning and artificial intelligence).

[26] *Id*.

platforms that host foundation models (among other resources, such as data). Developers can often download a foundation model, with all its statistical detail (e.g., parameters, weights) from a model hub and use it—or create a locally modified version of that foundation model—to create an application. With APIs, however, applications developers are able to programmatically communicate with models that may not otherwise be available for public use. That is, the only way to access a proprietary foundation model is through its API. Firms in the model layer operate in three primary ways: some firms are fully integrated, having their own proprietary data, models, and APIs, which are used to develop proprietary applications; some firms compile data into models and make those models available via model hubs or APIs, thereby creating room for downstream application development; and some firms are more disaggregated, offering, for example, discrete data services or serving only as a model hub.

Finally, we conclude Part I with a discussion of the application layer. Applications are the part of the sector that consumers interact with most directly: When we ask ChatGPT to tell us a joke about AI,[27] we use an application (ChatGPT). The application draws on all prior layers in the stack: it interacts with a model (GPT4); that model is stored in a cloud computing platform (Microsoft's Azure); and that platform requires microprocessing hardware (designed by Nvidia and fabricated by TSMC).

With this deeper and clearer understanding of the AI technology stack, we turn in Part II to the economic, national security, social, and political problems that currently exist or seem likely to emerge from the concentrated market structure within and across layers in the AI technology stack. We focus first on the traditional subjects of competition law and policy—extractive prices, quality of service concerns, self-preferencing and other forms of discrimination, and harms to downstream innovation, among other concerns.[28] Concentration at critical points in the AI technology stack also raises important national security and resilience concerns. If elements of production are limited to a single company or location, their failure could have significant ramifications for critical infrastructure—and for the economy more broadly.[29] Concentration in the AI technology sector is likely, moreover, to exacerbate concerns about economic inequality across society. Concentration can not only lead to a small number of firms with outsized economic power, it can also concentrate wealth in a small number of individuals—executives and shareholders.[30] Finally, the market structure at and across layers in the AI technology stack is concerning for the future of democracy. Concentration in AI may give a relatively small number of companies an outsized influence over the information ecosystem, complementing the outsized political influence they gain from their growing wealth and power.[31]

Our analysis of the AI technology stack and the downsides of an AI oligopoly yield four important conclusions, particularly in view of some of the

[27] One of us tried it. The joke was not funny.

[28] *See infra* Part II.A.

[29] *See infra* Part II.B.

[30] *See infra* Part II.C.

[31] *See infra* Part II.D.

prevailing tropes regarding regulating artificial intelligence. First, some commentators have worried that AI is moving too quickly for regulation.[32] We disagree. Even as technologists make rapid advances in AI technology, and as AI applications spread quickly across the economy, our analysis shows that the fundamentals of the technology and the basic industrial organization of the supply chain seem relatively stable. Many harms are thus already identifiable and are independent of improvements in the quality of AI applications, processing power, or other product developments. Moreover, as we note, a wait-and-see approach may make no-regulation, or weak regulation, a more likely scenario as it provides time for AI companies to entrench their power in the economy and politics.

Second, and relatedly, even as many of the harms we describe (though not all) are the traditional subjects of antitrust law, antitrust enforcement is unlikely to be sufficient.[33] As we show, the AI tech stack is already severely concentrated at many layers. Because antitrust enforcement operates *ex post* and on a case-by-case basis, it could take years for cases to make it through the courts to address anticompetitive behaviors—and then, only in a one-off fashion. In the courts, many of the most relevant antitrust doctrines have been narrowed over the last forty years, rendering underenforcement more likely in a sector that seems structurally inclined towards consolidation and concentration. Such underenforcement presents the risk that anticompetitive behaviors will persist and that market power, distributional, security, and democratic concerns will become more acute. It is therefore essential, we argue, to draw on antimonopoly tools that operate ex ante—including industrial policy, the tools of networks, platforms, and utilities (NPU) law, public options, cooperative governance. These market shaping tools can help to prevent and avoid harms by shaping market structure and firm operations in advance.

Third, our analysis of the industrial organization of the AI sector shows, perhaps counterintuitively, that the current non-regulatory path is likely to hamper innovation—and that antimonopoly governance rules could encourage innovation.[34] This is contrary to the popular cliché that regulation hinders innovation. Vertical integration across the AI tech stack is likely to restrict the number of providers of services at downstream layers in the stack, reducing innovation and choice. Many antimonopoly tools, however, can create a level playing field for downstream businesses that rely on some foundational service. Hence, these tools have the effect of expanding commerce and innovation by reducing bottlenecks and concomitant anticompetitive conduct.

Finally, we suggest that governing market structure is critical to addressing many common concerns about the conduct of AI applications, including algorithmic bias as well as false or misleading AI determinations. This is for two reasons. First, many of these harms may themselves be derivative of market structure and market power: Market concentration and vertical integration can lead to fewer downstream applications. Greater competition, by contrast, may give rise to an AI

[32] *See* Part III.A. For one example, Amy Gibbons, *AI is Moving Too Fast to Regulate, Security Minister Warns*, THE TELEGRAPH, June 9, 2023, 9:11pm, https://www.telegraph.co.uk/news/2023/06/09/security-minister-artificial-intelligence-regulation/.

[33] See Part III.B.

[34] See Part III.C.

marketplace that includes, for example, less-biased or more privacy-protecting technologies—applications that may be more likely to win consumer approval. Second, a clear understanding of the sector's industrial organization helps clarify who to regulate and how to regulate them. Consider, for example, harms stemming from algorithmic bias: Even as policymakers have focused attention on biased applications, regulations might be better targeted at companies lower in the stack—in, say, the model layer—to address concerns about bias. This is so even if those companies only offer services in those lower layers, and do not develop AI applications at all. Clarity about industrial organization can therefore bring a great deal of specificity to the question of how to regulate AI.

In Part IV, we turn to more specific solutions. We outline how antimonopoly and competition tools—industrial policy, networks, platforms, and utilities (NPU) law, public options, and cooperative governance—can apply to the AI sector. In the hardware layer, for example, policymakers have already adopted industrial policies[35] to address scarcity and supply chain fragility in the production of semiconductors. We agree with this approach, particularly insofar as it is aimed at concerns about resilience and national security. But we also caution that industrial policy can and should be attentive to industrial organization, either by enhancing competition where feasible, or else by addressing the power of dominant firms. Second, the law of networks, platforms, and utilities (NPU law), has long governed sectors with tendencies toward monopoly and oligopoly. We show how various tools for regulating networks, platforms, and utilities (or NPUs)—structural separation requirements; nondiscrimination rules and open access requirements; interoperability mandates, as well as service and rate regulations—could be applied to the various layers in AI's technology stack. Third, we argue that public options[36] may be helpful complements at a number of places in the AI technology stack. Public provision of certain resources would increase competition, set an effective price floor, and ensure an open access baseline, all while providing a utility-like resource that can foster downstream innovation. Fourth, we discuss cooperative governance as one way to manage AI-related businesses. Cooperatives are firms in which users are owners. Historically, they have operated both as an antimonopoly tool and as a way to more equitably distribute the wealth of productive enterprises. So far as we are aware, our account is the first to consider the application of many of these tools to AI.[37]

[35] For purposes of this Article, we take a narrow definition of industrial policy, meaning investments to spur domestic industrial production in a particular sector. For discussions that make the case for a broader definition of the term, see Todd Tucker, *Industrial Policy and Planning: What It Is and How to Do It Better*, ROOSEVELT INSTITUTE (July 30, 2019); Ganesh Sitaraman, *Industrial Revolutionaries*, AM. PROSPECT (Sept. 10, 2020).

[36] GANESH SITARAMAN & ANNE ALSTOTT, THE PUBLIC OPTION (2019).

[37] So far as we are aware, there has not been any sustained work applying public utilities tools to AI specifically. There has been discussion of concentration in the cloud layer, but not framed around AI. See House Digital Markets report. A few commentators have suggested a public option for AI, but in popular writings and without much analysis. *See* Ben Gansky, Michael Martin & Ganesh Sitaraman, *Artificial Intelligence is Too Important to Leave to Google and Facebook Alone*, N.Y. TIMES, Nov. 10, 2019; Bruce Schneier & Nathan E. Sanders, *Build AI by the People, for the People*, FOREIGN POL'Y (June 12, 2023, 10:34 AM),

In arguing for an antimonopoly approach to governing AI, we make four contributions. First and most directly, we show that serious market power and competition problems already exist—and are likely to persist—in the AI technology stack, and we describe how policymakers can address them. These concerns have received comparatively little attention in the scholarly and policy debates over regulating AI. Second, we make the case that antitrust enforcement is likely to be insufficient for governing AI's market structure problems and advocate for a renewed focus on affirmative forms of regulation and governance. Third, for those who are primarily interested in the *uses* of AI, rather than its market structure, our account of the industrial organization of the AI sector offers a helpful framework for policy development. Identifying the specific layers and sublayers of the AI stack should inform the design of regulations that seek to address the uses and abuses of AI. Finally, and more broadly, our work contributes to the recent revival of NPU law,[38] and indirectly to the law and political economy (LPE) movement.[39] NPU law has, until recently, lain fallow, with its legal tools largely disappearing from policy debates.[40] We show here how its tools can be extremely useful in governing the emergence of a frontier technology. In doing so, we also align with the LPE movement's broader attention to political economy, rather than a more limited focus on economic efficiency, and we show that concentration in the AI sector has implications for national security and resilience, distribution, and for democracy. Public policy must contend with questions beyond economic analysis, including the vast power and distributional concerns that might emerge from control of this technology.[41]

A few clarifications are also in order. First and foremost, we do not aim to address every potential problem with AI[42] or to provide a comprehensive approach to AI governance. Our focus here is on market concentration and its harms Second, while we show how antimonopoly tools can operate at different layers in the AI stack, we do not address the best way to adopt these tools. Some NPU tools could likely be applied via the common law,[43] or through notice and comment

https://foreignpolicy.com/2023/06/12/ai-regulation-technology-us-china-eu-governance/. There is far more on antitrust enforcement, but again, this is usually framed around technology platforms generally, rather than AI specifically.

[38] MORGAN RICKS, GANESH SITARAMAN, SHELLEY WELTON & LEV MENAND, NETWORKS, PLATFORMS, AND UTILITIES: LAW AND POLICY (2022). For an application of this body to tech platforms, see K. Sabeel Rahman, *The New Utilities: Private Power, Social Infrastructure, and the Revival of the Public Utility Concept*, 39 CARDOZO L. REV. 1621 (2018).

[39] Jedediah Britton-Purdy et al., *Building a Law and Political Economy Framework*, 129 YALE L.J. 1784 (2020)

[40] For an account of the abandonment of NPU tools across sectors, see Joseph D. Kearney & Thomas W. Merrill, *The Great Transformation of Regulated Industries Law*, 98 COLUM. L. REV. 1323 (1998).

[41] *See* DARON ACEMOGLU & SIMON JOHNSON, POWER AND PROGRESS (2023).

[42] For a helpful overview of many of the downstream, application-based problems, see Laura Weidinger et al., *Taxonomy of Risks Posed by Language Models*, FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (June 2022), https://dl.acm.org/doi/10.1145/3531146.3533088.

[43] Ganesh Sitaraman & Morgan Ricks, *Tech Platforms and the Common Law of Carriers*, DUKE L.J. (forthcoming 2024).

rulemaking under current law.[44] Competition laws are already on the books and industrial policy efforts are currently underway. And any of these tools could be adopted (or adapted) by statute. Whatever the pathway for implementation, our ultimate hope is that this Article helps build the case for a different vision of a world with artificial intelligence, one in which the public has more control over and say in the future of this critical technology.

I. UNDERSTANDING THE AI TECHNOLOGY STACK

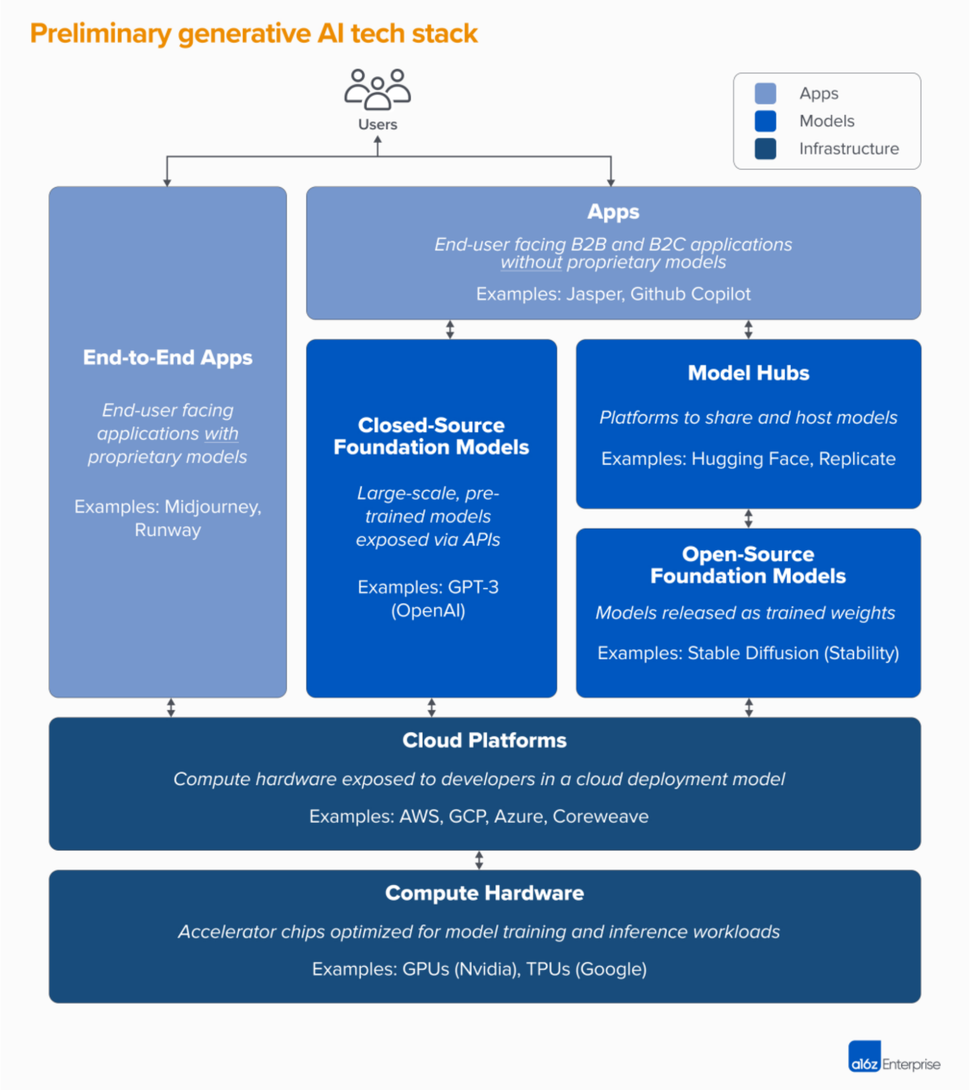
Policymaking requires understanding the technologies and industries at issue. For all the discussions of regulating AI, in this Part, we offer what we believe is the first account in the legal literature of AI's technology stack—the industrial and technological organization of AI.[45] AI's technology stack consists of four major layers, with some containing nested sublayers. The first layer consists of hardware—predominantly microchips that provide processing power. The second layer is cloud computing, which includes infrastructural capacity (e.g., data storage, processing capacity, and network connectivity), alongside related services. Three sublayers comprise the third layer: data; models trained on that data; and modes of accessing those models (and their underlying data), predominantly through hubs or application programming interfaces (APIs). The final layer consists of applications—the layer through which most consumers interact with AI.

At each layer, we provide an overview of the layer, its components and uses, and its market structure. This forms the foundation for identifying where policy problems are likely to emerge—and how to address them.[46]

[44] *See* Spiro, *supra* note xx.

[45] Some scholars have described parts of this stack. *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn about Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017). But we believe that our account is the first comprehensive assessment of the market structure of the entire stack, from microprocessing hardware to applications.

[46] Our account of these layers aligns with a number of accounts from technology industry analysts. *See, e.g.*, Matt Bornstein, Guido Appenzeller & Martin Casado, *Who Owns the Generative AI Platform?*, ANDREESENHOROWITZ (Jan. 19, 2023), https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform/; Brad Smith, *Governing AI: A Blueprint for Our Future*, TOOLS AND WEAPONS WITH BRAD SMITH (May 30, 2023), https://tools-and-weapons-with-brad-smith.simplecast.com/episodes/governing-ai-a-blueprint-for-our-future/transcript; Sayash Kapur & Arvind Narayanan, *Three Ideas for Regulating Generative AI*, AI SNAKE OIL, June 21, 2023, https://aisnakeoil.substack.com/p/three-ideas-for-regulating-generative; Matt McIlwain, *Game On in the Generative AI Stack*, MADRONA, Mar. 20, 2023, https://www.madrona.com/game-on-in-the-generative-ai-stack/; Deedy, @debarghya_das, TWITTER, Mar. 16, 2023, 8:45pm, https://twitter.com/debarghya_das/status/1636544140069711872. *See also* Assaf Araki, *Demystifying the AI Infrastructure Stack*, INTEL CAPITAL, April 2, 2020, https://www.intelcapital.com/demystifying-the-ai-infrastructure-stack/;


Preliminary generative AI tech stack
Users
Apps
Models
Infrastructure
Apps
End-user facing B2B and B2C applications without proprietary models
Examples: Jasper, Github Copilot
End-to-End Apps
End-user facing applications with proprietary models
Examples: Midjourney, Runway
Closed-Source Foundation Models
Large-scale, pre-trained models exposed via APIs
Examples: GPT-3 (OpenAI)
Model Hubs
Platforms to share and host models
Examples: Hugging Face, Replicate
Open-Source Foundation Models
Models released as trained weights
Examples: Stable Diffusion (Stability)
Cloud Platforms
Compute hardware exposed to developers in a cloud deployment model
Examples: AWS, GCP, Azure, Coreweave
Compute Hardware
Accelerator chips optimized for model training and inference workloads
Examples: GPUs (Nvidia), TPUs (Google)
a16z Enterprise

Source: Andreessen-Horowitz[47]

A. Hardware

The technological foundation of AI is computer hardware—specifically, computer microprocessing units (or, colloquially, chips) that try to "pack [in] the maximum number of transistors" to quickly make the enormous number of calculations required by AI.[48] Chips come in three basic varieties. The primary of these are GPUs (graphical processing units), which were originally designed for processing images—a task that benefits from parallel (rather than sequential)

[47] Matt Bornstein, Guido Appenzeller & Martin Casado, *Who Owns the Generative AI Platform?*, ANDREESENHOROWITZ (Jan. 19, 2023), https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform/

[48] Saif M. Khan & Alexander Mann, *AI Chips: What They Are and Why They Matter*, CENTER FOR SECURITY AND EMERGING TECHNOLOGY 3 (April 2020)

computation.[49] But because that is true for more than just image processing—including for AI—GPUs are now general purpose chips, and have become dominant for training AI models.[50] The other two main types of microchips are field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs). ASICs are notable because they are, as their name suggests, application specific: The chips are optimized to run specific tasks, which could be helpful for deploying certain AI applications at scale.[51]

The amount of processing power—and therefore the number of microprocessors—needed for AI is extraordinary. As we describe in more detail below, AI models first need to be "trained," meaning that a specific AI algorithm is initially developed and refined. Then a trained algorithm works through "prediction" or "inference," in which the algorithm is deployed to engage a real-world scenario in light of its training. According to some estimates, "hundreds of GPUs are required to train artificial intelligence models," and eight GPUs might be required to respond to a single query on Bing's search using ChatGPT.[52] Companies that seek to deploy AI at scale thus need a significant amount of computing power. Meta, for example, used $25 million worth of Nvidia A100 chips (released in 2020), alongside $100,000 in electrical and power consumption costs, to train its LLaMA-65B model.[53] Microsoft might need more than 20,000 GPU servers with 8 chips each to operate ChatGPT for all Bing users.[54] At a price of $10,000 a chip for the Nvidia's A100, or $200,000 for its 8-chip system, the cost of deploying AI at scale is huge. For Google, which answers many more queries per day than Bing, some estimate the cost could be $80 billion dollars.[55]

Each new generation of GPU accelerates AI development, because microchips of prior generations seem increasingly "larger, slower, and more power hungry" and thus give rise to "huge energy consumption costs" that are "unaffordable" for all but some of the largest and most well-capitalized firms.[56] Nvidia's newer chip, the H100, was released in 2022 at a cost of $40,000.[57] Its performance is estimated to be three times better than its previous model.[58] Google has already

[49] *Id*. at 18.

[50] *Id*.

[51] *Id*. at 20-21.

[52] Kif Leswing, *Meet the $10,000 Nvidia Chip Powering the Race for A.I.*, CNBC, Feb. 23, 2023, https://www.cnbc.com/2023/02/23/nvidias-a100-is-the-10000-chip-powering-the-race-for-ai-.html

[53] Joe Lamming, *GPT-4: The Giant AI (LLaMA) is Already Out of the Bag*, VERDANTIX, April 5, 2023, https://www.verdantix.com/insights/blogs/gpt-4-the-giant-ai-llama-is-already-out-of-the-bag

[54] Leswing, *supra* note xx.

[55] *Id*.

[56] Khan & Mann, *supra* note xx, at 6.

[57] Tim Bradshaw & Richard Waters, *How Nvidia Created the Chip Powering the Generative AI Boom*, FIN. TIMES, May 26, 2023, https://www.ft.com/content/315d804a-6ce1-4fb7-a86a-1fa222b77266

[58] Id.

built a supercomputer with 26,000 of the new GPUs.[59] Given the high energy costs, large technology companies often choose to physically locate their data operations close to sources of cheap electricity.[60]

The structure of the market for microprocessors is highly concentrated. As chip technologies have become more and more sophisticated, fewer firms are able to supply the needed technologies. While reports differ, Nvidia—which designs chips—appears to have captured between 80 and 95 percent market share of the GPU chip business used for AI.[61] Nvidia's chips are, in turn, manufactured (or "fabricated") by Taiwan Semiconductor Manufacturing Corporation (TSMC),[62] which is far and away the dominant semiconductor manufacturer.[63] Only Samsung also fabricates the smallest, highest powered chips.[64] To make the smallest chips requires photolithography equipment, something only one company in the world, the Dutch firm ASML,[65] provides—and sells for between $150 and $200 million per machine.[66]

[59] Kyle Wiggers, *Meta Bets Big on AI with Custom Chips—and a Supercomputer*, TECHCRUNCH, May 18, 2023, https://techcrunch.com/2023/05/18/meta-bets-big-on-ai-with-custom-chips-and-a-supercomputer/

[60] KATE CRAWFORD, ATLAS OF AI 215–216 (2021).

[61] Leswing, *supra* note xx (noting that Nvidia has 95% market share for machine learning); Zoe Corbyn & Ben Morris, *Nvidia: The Chip Maker that Became and AI Superpower*, BBC NEWS, May 30, 2023, https://www.bbc.com/news/business-65675027 (same); Asa Fitch & Jiyoung Sohn, *The Next Challengers Joining Nvidia in the AI Chip Revolution*, WALL ST. J., July 11, 2023 (12:00am ET), https://www.wsj.com/articles/the-next-challengers-joining-nvidia-in-the-ai-chip-revolution-e0055485 (noting that Nvidia has more than 80% of the market); Wallace Witkowski, *Nvidia 'Should Have At Least 90%' of AI Chip Market with AMD on its Heels,* MARKETWATCH, July 10, 2023 (12:50pm ET), https://www.marketwatch.com/story/nvidia-should-have-at-least-90-of-ai-chip-market-with-amd-on-its-heels-13d00bff (projecting Nvidia will have 90% of the chip market).

[62] Arjun Kharpal, *Two of the World's Most Critical Chip Firms Rally After Nvidia's 26% Share Price Surge*, CNBC, May 25, 2023 (7:14 AM EDT), https://www.cnbc.com/2023/05/25/tsmc-asml-two-critical-chip-firms-rally-after-nvidias-earnings.html

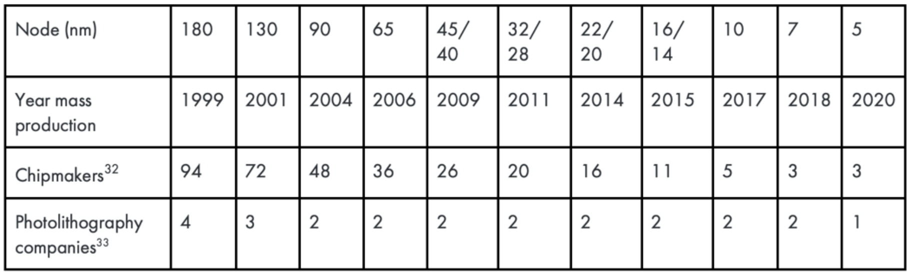
[63] TSMC's market share was estimated at 58.5% in 2022, with runner-up Samsung coming in at 15.8%. Peter Clarke, *TSMC, Globalfoundries Gained as Foundry Market Cooled*, EENEWS, March 13, 2023, https://www.eenewseurope.com/en/tsmc-globalfoundries-gained-as-foundry-market-cooled/.

[64] Khan & Mann, *supra* note xx, at 11.

[65] *Id*.

[66] Kate Tarasov, *ASML is the Only Company Making the $200 Million Machines Needed to Print Every Advanced Microchip. Here's an Inside Look*, CNBC (Mar. 23, 2022, 1:00pm EDT), https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html.

Table 1: Number of companies at each node



Node (nm)	180	130	90	65	45/ 40	32/ 28	22/ 20	16/ 14	10	7	5
Year mass production	1999	2001	2004	2006	2009	2011	2014	2015	2017	2018	2020
Chipmakers[32]	94	72	48	36	26	20	16	11	5	3	3
Photolithography companies[33]	4	3	2	2	2	2	2	2	2	2	1

Source: Khan & Mann, *supra* note xx, at 12.

In recent years, other companies—large technology platform companies—have moved towards the chip design business. Meta, for example, has designed a chip specific for certain training and inference functions. A Meta executive explained that "Building our own [hardware] capabilities gives us control at every layer of the stack, from datacenter design to training frameworks….This level of vertical integration is needed to push the boundaries of AI research at scale."[67] Google, Amazon, and Microsoft have all likewise developed their own respective chips, designed for specific AI-related functions.[68] Some of these, like Google's "Tensor Processing Unit," or TPU are not general purpose GPUs, but ASICs.[69] These chips may be particularly useful for deploying inferential capabilities at scale, because they can be designed to make specific tasks especially fast. But such specialization also means reduced flexibility to execute other workloads or to change as AI applications are updated.[70]

B. *Cloud Infrastructure*

AI's capabilities arise out of two massively scaled resources: data and computing power. Developers "train" AI models on enormous quantities of data until deciding that the model is ready to be deployed. As we have noted, training (and, eventually, inference) require significant processing power—sometimes called computationally capacity or "compute"—in order to complete the substantial number of calculations needed to develop a model and provide "intelligent" responses. To reach the necessary scale of compute, providers have relied on cloud infrastructure.

[67] Kyle Wiggers, *Meta Bets Big on AI with Custom Chips—and a Supercomputer*, TECHCRUNCH, May 18, 2023, https://techcrunch.com/2023/05/18/meta-bets-big-on-ai-with-custom-chips-and-a-supercomputer/

[68] Google's is called the Tensor Processing Unit (TPU). AWS (Amazon) has Tranium and Inferencia, and Microsoft is developing Athena, in conjunction with chip company AMD. *Id.*

[69] Nicole Kobie, *Nvidia and the Battle for the Future of AI Chips*, WIRED, June 17, 2021 (10:10AM), https://www.wired.co.uk/article/nvidia-ai-chips

[70] Khan & Mann, *supra* note xx, at 20-21.

In general, cloud computing refers to the "ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources."[71] The "cloud" is simply hardware that exists somewhere else: It is a set of computers, servers, storage, cables, and other hardware that are typically concentrated in gigantic warehouses and to which users connect remotely (e.g., over the internet).

These hardware resources are used to offer three general categories of services: software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (SaaS).[72] Software as a service is the most familiar to the average consumer: It is the ability to run an application on one's own device, even as the application works by connecting to the provider's remote servers or networks.[73] Google Docs is one exemplar. Platform as a service is more relevant to developers: It allows a user to connect to the remote infrastructure in order to use providers' "programming languages, libraries, services, and tools."[74] Infrastructure as a service provides users with "processing, storage, networks, and other fundamental computing resources."[75] While all three categories of cloud computing are relevant to AI, we focus here primarily on infrastructure as a service, because both AI models and applications rely on these infrastructural capabilities.

Cloud infrastructure features several dynamics that tend toward concentration and make sustaining competition difficult. First are extremely high capital costs. Building data centers, server farms, and the networked systems to connect them is expensive. Some have described the cost as "bigger than building a cellular network" and as within reach only "for countries and major companies."[76] Second, there are significant switching costs to moving from one provider to another. Some of these switching costs are inherent to provider variation: Customers might need to change aspects of their business, and hire developers who can work across multiple platforms, or else risk disrupting reliable, continuous, and seamless service to their own consumers.[77] Some businesses have declared that even a 20 percent price discount is insufficient to overcome these concerns.[78] Such impediments to switching are exacerbated by additional costs imposed by cloud computing providers, who sometimes charge "egress fees"—akin to termination fees—that charge

[71] Peter Mell & Timothy Grace, *The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology* 6, NAT. INST. STANDARDS & TECH. (2011) https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

[72] Id. at 2-3.

[73] Id.

[74] *Id*.

[75] *Id*. at 3.

[76] Investigation of Competition in Digital Markets: H. Comm. On The Judiciary, 117th Congress, 96 (2020) https://www.govinfo.gov/content/pkg/CPRT-117HPRT47832/pdf/CPRT-117HPRT47832.pdf

[77] *Id*. at 269.

[78] Kamila Benzina, *Cloud Infrastructure-As-A-Service as an Essential Facility: Market Structure, Competition, and the Need for Industry and Regulatory Solutions*, 34 BERKELEY TECH. L.J. 119, 133 (2019).

customers for taking their data out of one cloud computing system and to another.[79] Third, cloud computing systems are subject to network effects: the more users on a single cloud system, the more developers will make applications designed for that cloud system—which, in turn, attracts more users.[80] This problem is made more difficult because developers may build expertise in operating in one cloud system, making it more likely a firm will adopt a dominant cloud provider.[81]

Given these dynamics, the market structure of cloud computing has consolidated among three primary businesses: Amazon Web Services (or AWS), Microsoft Azure, and Google Cloud Platform.[82] The specific market shares of the firms vary by year and analyst but are remarkably consistent. AWS is far and away the dominant provider, with more than 30 percent market share—and approaching 40 percent in some assessments. Azure comes in second and is near 20 percent, and Google and others run further behind.[83] Hence, the market for cloud computing service is characterized by oligopoly. "With identical services comes commoditization, and only big vendors that can deliver huge economies of scale with margins will survive in this space."[84] As a result, commentators have, as early as 1961, analogized cloud computing to other basic utilities.[85]

[79] On egress fees, see Investigation into Digital Markets, *supra* note xx, at 269. For a discussion of the high costs of compute power, see Guido Appenzeller, Matt Bornstein & Martin Casado, *Navigating the High Cost of AI Compute*, ANDREESENHOROWITZ, April 27, 2023, https://a16z.com/2023/04/27/navigating-the-high-cost-of-ai-compute/.

[80] *Id*. at 97

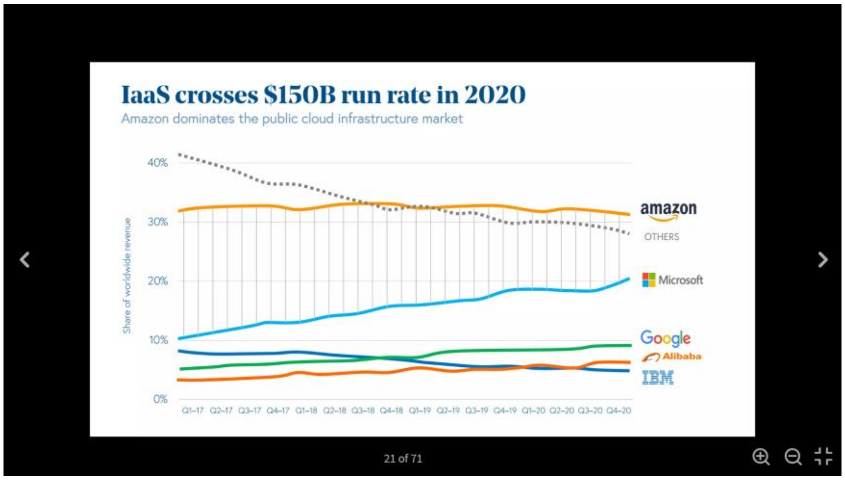
[81] *Id*. at 269.

[82] Investigation into Digital Markets, *supra* note xx, at 92.

[83] *See, e.g.*, Gartner Says Worldwide IaaS Public Cloud Services Market Grew 41.4% in 2021, GARTNER, https://www.gartner.com/en/newsroom/press-releases/2022-06-02-gartner-says-worldwide-iaas-public-cloud-services-market-grew-41-percent-in-2021 (June 2, 2022); STATISTA, United States Cloud Infrastructure Services Vendor Market Share Q1 2021 (Dec. 6, 2022).

[84] McKendrick, *supra* note xx.

[85] John McCarthy, speaking at the MIT Centennial (1961), *in* SIMSON L. GARFINKEL, ARCHITECTS OF THE INFORMATION SOCIETY, THIRTY-FIVE YEARS OF THE LABORATORY FOR COMPUTER SCIENCE at 1, § 1 (1999) ("[C]omputation may someday be organized as a public utility, just as the telephone system is a public utility. We can envisage computer service companies whose subscribers are connected to them .... Each subscriber needs to pay only for the capacity that he actually uses. . . ."); Bob O'Donnell, *Cloud Computing As A Utility Is Going Mainstream*, VOX (Aug. 17, 2016) https://www.vox.com/2016/8/17/12519046/cloud-computing-as-utility-private-public-data-center; Rod Paddock, *The Cloud Networking Effect*, CODE MAGAZINE (Dec. 16, 2021) https://www.codemag.com/article/1301011/The-Cloud-Networking-Effect ("You wouldn't set up your own gas fired power plant to supply power to your home. So why would you bother setting up your own server infrastructure?").



IaaS crosses $150B run rate in 2020

Amazon dominates the public cloud infrastructure market

Source: Bessemer Venture Partners[86]

C. The Model Layer

Once providers secure access to hardware and other infrastructural requirements—processing power, storage, bandwidth, and computational capacity—they can turn, finally, to developing the "intelligence" in artificial intelligence. Such intelligence rests upon a statistical model for completing whatever tasks will eventually be assigned to that AI application.

Consider, for example, applications of large language models, which are used today to generate novel text. Large language models, such as ChatGPT, begin with extremely large corpuses of text. GPT-3, for example, is based on 300 billion "tokens" of text,[87] sampled from nearly 500 billion of such tokens extracted from a range of sources, including over a decade of internet text, fifteen years of Reddit posts, two online repositories of books, and English-language Wikipedia.[88] GPT-4 depends on even more training data, though OpenAI has been less forthcoming about the sources and quantities of training data it depends upon, except to say that this newest iteration expands upon the resources used to develop GPT-3.[89]

[86] BESSEMER VENTURE PARTNERS, "State Of The Cloud 2021" (March 10, 2021) https://www.bvp.com/atlas/state-of-the-cloud-2021

[87] A token is a computational representation of text (ranging from about four characters in the case of OpenAI's systems, see https://platform.openai.com/tokenizer, to as much as common two- or three-word phrases).

[88] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei, *Language Models Are Few-Shot Learners*, PROC. 34TH CONF. NEURAL INFO. PROCESSING SYS. (2020) (introducing and describing the GPT-3 language model)

[89] OPEN AI, GPT-4 TECHNICAL REPORT 2 (2023), https://arxiv.org/pdf/2303.08774.pdf ("Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training

This data forms the basis for the model, which is, in simplified terms, little more than a statistical representation of all the input data. In the case of large language models, for example, the model is "trained" to "understand" that, the text following "Jack" is more probabilistically likely to be "and Jill" or "of spades"—and not, say, "and Heather" or "of rakes." And such assessments are made on continuous basis: seeing "Jack and" increases the likelihood of seeing "Jill" or "Diane" next; "Jack of" increases the likelihood of seeing "all trades" or "spades." This continuous representation of the relationship among tokens (i.e., snippets of text) and sets of tokens, comprise the model.[90] These basic, or "foundation," models, may, moreover, be tweaked or "fine-tuned" to particular purposes or applications.[91]

Downstream developers need to access the foundation models for fine-tuning, and for use in a particular application (ChatGPT to generate text for a customer service chatbot, for example). Some models, like OpenAI's, are closely held by their developer, and are accessible only via an API (or application programming interface). An API is "tool that allows programmers to use pre-written code to build certain functions into their own programs,"[92] or, put more simply, an API provides "the necessary infrastructure for [downstream] computer programmers to develop new programs and applications" that build upon the model.[93] The developer of a chatbot program might access GPT-3 through an API, using various commands to "fine-tune" the model for specific purposes,[94] and then to send prompts to GPT and retrieve responses.[95] Other foundation models and their final statistical weights and measures are open source. LLaMA 2, which was developed by Meta, is an example. Open source models are hosted on public websites, known as "model hubs" for others to download and use. Model hubs, such as Hugging Face, host models, the underlying data, and APIs all for use by developers.

The model layer itself therefore consists of several layers, and this multi-layer structure has important implications for market structure and competition—some of which are entangled with the concerns raised above regarding hardware and computational capacity.

First is the data layer. As noted above, developing a model often depends on access to vast troves of data. In some instances, such data may be comparatively cheap to obtain. The repositories of internet text and Reddit posts used by

compute, dataset construction, training method, or similar."). Some reports have put the training of GPT-3 at $10-12 million for each training run. Alex Hern, *TechScape: Will Meta's Massive Leak Democratise AI—and at What Cost?*, THE GUARDIAN (March 7, 2023, 06:45 EST).

[90] For a more technical discussion of how LLMs work, see Stephan Wolfram, *What is ChatGPT Doing…and Why Does it Work?*, Feb. 14, 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

[91] See Rishi Bomnasani et al., *On the Opportunities and Risks of Foundation Models*, Center for Research on Foundation Models, Stanford Institute for Human-Centered Artificial Intelligence (HAI), at https://arxiv.org/abs/2108.07258 (manuscript)

[92] Google LLC v. Oracle America, 593 U.S. __ (2021).

[93] *Id*.

[94] OpenAI, *Fine Tuning*, OPENAI, https://platform.openai.com/docs/guides/fine-tuning.

[95] OpenAI, *GPT Models*, OPENAI, https://platform.openai.com/docs/guides/gpt.

ChatGPT, for example, are freely available for download.[96] Other providers have similarly scraped millions of publicly available images from online sources to train facial-recognition models.[97] But data can also be difficult and expensive to obtain, and developers may place a premium on exclusive access to important sources of training data.[98] Google, for example, struck a controversial deal with Ascension Health Systems for access to patient records for the purposes of training diagnostic and other medical AI systems.[99] Google also paid $2.1 billion to purchase Fitbit, again with an eye towards collecting health metrics and data.[100] Importantly, data may not be immediately usable for training an AI model. Data, which is sometimes stored in unstructured "data lakes," often requires some combination of cleaning, validation, transformation, and labeling before it can be used for model training.

Second is the model layer. As noted above, training a model is often computationally intensive—meaning that it can be hugely expensive—depending on the nature of the algorithm.[101] As training becomes more computationally complex, requirements in the processing and hardware layers increase dramatically because electrical power, processing, and storage and bandwidth requirements can grow polynomially (or even exponentially). These barriers to entry are significant: Developing a foundation model often requires a substantial capital investment.[102] Some commentators have suggested that OpenAI was able to successfully develop GPT-3 only because it was a "a well-capitalized company" that also "teamed up with Microsoft to develop an AI supercomputer," but that similar successes seem "potentially beyond the reach of [other] AI startups . . . which in some cases lack the capital required."[103] As a result, the number of foundation models that can (or

[96] COMMON CRAWL, *About*, https://commoncrawl.org/about/; OPENWEBTEXT2, Welcome!, https://openwebtext2.readthedocs.io/en/latest/; *see also* Brown et al., *supra* note xx, (describing the training data for GPT-3).

[97] Olivia Solon, *Facial Recognition's 'Dirty Little Secret': Millions of Online Photos Scraped Without Consent*, NBC NEWS (Mar. 17, 2019, 10:25 AM), https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921 [https://perma.cc/9WNY-YR8A]; Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It,* N.Y. TIMES (Nov. 2 , 2021), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facialrecognition.html [https://perma.cc/4LFZ-4GSD].

[98] See, e.g., Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018); C. Scott Hemphill, *Disruptive Incumbents: Platform Competition in an Age of Machine Learning*, 119 COLUM. L. REV. 1973, 1978–79 (2019).

[99] Rob Copeland, *Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans*, WALL ST. J. (Nov. 11, 2019, 4:27 PM),

[100] *Id.*

[101] *See* Part IA above, for discussions of the compute costs and chip and electricity needs. For another account of the costs of training machine learning systems, consider Ben Cottier, *Trends in the Dollar Training Cost of Machine Learning Systems*, EPOCH, Jan. 31, 2023, https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.

[102] Kyle Wiggers, *OpenAI's Massive GPT-3 Model Is Impressive, But Size Isn't Everything*, VENTURE BEAT (June 1, 2020, 1:05 PM), https://venturebeat.com/2020/06/01/ai-machinelearning-openai-gpt-3-size-isnt-everything

[103] *Id.*

that perhaps should, from the standpoint of productive efficiency) exist for any given application class—language, image generation—may be quite limited.[104]

The final layer regards access to the trained model. Foundation model developers can choose the terms on which they make that model available to the public (if at all). Though LLaMA was likely very expensive to develop, Meta and Microsoft, for example, have decided to make that model public at no cost. Why? They are likely betting that providing an such an open platform will stimulate the development of downstream applications in a way that redounds to their ultimate benefit, much as Microsoft has long encouraged the development of third-party applications that could run on its Windows platform.[105] This is in part because the cost of finetuning a model is miniscule compared to the costs of training it.[106] By contrast, OpenAI (which, again, counts Microsoft as a significant investor) has declined to open-source its GPT models—though it does allow third-party developers to build upon those models through the APIs that it develops, documents, and makes available to the public.[107]

Notably, a number of the biggest technology companies participate at every stage in the model layer (and, as we have seen, in earlier layers too). Google, for example, is developing its own chips (TPUs), has its own cloud infrastructure (Google Cloud), collects enormous amounts of data, has its own foundation models (PaLM 2, Codey, Imagen, and Chirp), and offers applications (such as Bard, a competitor to ChatGPT).[108] In other words, Google offers a vertically-integrated, closed-source artificial intelligence system all the way up and down the AI technology stack. Microsoft's massive investment into OpenAI has placed it in a

[104] This may be especially true if we consider the carbon costs of model "overbuilding." Emma Strubell, Ananya Ganesh & Andrew McCallum, *Energy and Policy Considerations for Deep Learning in NLP* (June 5, 2019) (unpublished manuscript), https://arxiv.org/pdf/1906.02243.pdf [https://perma.cc/2LH6-9YXV] (describing costs in terms of power and shared computing resource prices); Kate Saenko, *Feed Me, Seymour!—Why AI Is so Power-Hungry*, ARSTECHNICA (Dec. 29, 2020, 6:38 AM), https://arstechnica.com/science/2020/12/why-ai-is-so-power-hungry/?comments=1 [https://perma.cc/XB8S-QQU2] (citing the previous source and explaining that the power consumption demands of training and optimizing one machine-learning-based language model is equivalent to the cost of flying "315 passengers, or an entire 747 jet" on a "round trip between New York and San Francisco").

[105] Steve Inskeep & Olivia Hampton, *Meta Leans on 'Wisdom of Crowds' in AI Model Release*, NPR, July 19, 2023 5:11AM ET, https://www.npr.org/2023/07/19/1188543421/metas-nick-clegg-on-the-companys-decision-to-offer-ai-tech-as-open-source-softwa (quoting Nick Clegg as saying Meta is "not a charity" and that it made LLaMa "available for free to the vast majority of those who will use it" notwithstanding the fact that it was "an expensive endeavor to have built [LLaMa] in the first place" because doing so was "in [Meta's] interest," as it will "help set in motion a kind of flywheel of innovation which [Meta] can then incorporate into [its] own products.").

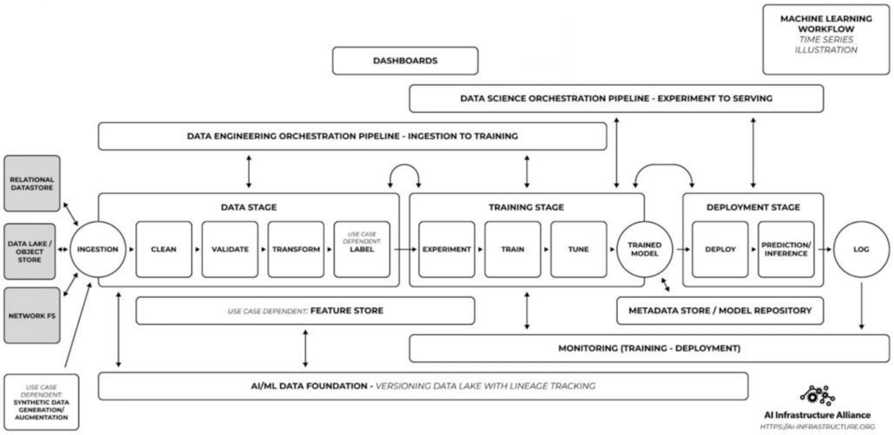
[106] *See* Dylan Patel & Afzal Ahmad, *Google "We Have No Moat, And Neither Does OpenAI,"* SEMIANALYSIS, May 4, 2023, https://www.semianalysis.com/p/google-we-have-no-moat-and-neither.

[107] For a discussion of how OpenAI is not so open, see Chloe Xiang, *OpenAI is Now Everything It Promised Not to Be: Corporate, Closed-Source, and For-Profit*, MOTHERBOARD: TECH BY VICE (Feb. 28, 2023, 10:35AM), https://www.vice.com/en/article/5d3naz/openai-is-now-everything-it-promised-not-to-be-corporate-closed-source-and-for-profit.

[108] Janakiram MSV, *Google's Generative AI Stack: An In-Depth Analysis*, THE NEW STACK, May 31, 2023 (7:59 AM), https://opendatascience.com/the-rapid-evolution-of-the-canonical-stack-for-machine-learning/

similar role: from chips and Azure cloud services to OpenAI's closed-source system of models and APIs and applications. Given the high barriers to entry across these layers, including aspects of the model layer, and the significant first-mover advantages in model development and application deployment, it is likely that these companies will develop and retain control over a significant share of this layer in the AI sector.

The availability of open source models is unlikely to upend this dominance. It is true that various resources may be available on open-source terms: structured and unstructured data (i.e., data warehouses and data lakes), certain models, and APIs. But even developers relying on these resources may need to depend on dominant cloud operators to achieve scale. As an interdisciplinary team of researchers recently explained, "while a handful of maximally open AI systems exist," "the resources needed to build AI from scratch, and to deploy large AI systems at scale, remain 'closed'—available only to those with significant (almost always corporate) resources."[109] That is, the availability of open source resources in the model layer does little to upend concerns about concentration in the "lower" layers—cloud computing and microprocessing. Moreover, even in the model layer, concerns about concentration persist. The best foundation models require enormous amounts of data—but usable data is not always freely and easily available. Training foundation models, as we have seen, also has extremely high compute costs, raising entry barriers. Hence, it is possible that high-quality data resources and associated models will concentrate into a small number of dominant players.


MACHINE LEARNING WORKFLOW
TIME SERIES ILLUSTRATION
DASHBOARDS
DATA SCIENCE ORCHESTRATION PIPELINE - EXPERIMENT TO SERVING
DATA ENGINEERING ORCHESTRATION PIPELINE - INGESTION TO TRAINING
RELATIONAL DATASTORE
DATA LAKE / OBJECT STORE
NETWORK FS
DATA STAGE
INGESTION
CLEAN
VALIDATE
TRANSFORM
USE CASE DEPENDENT: LABEL
TRAINING STAGE
EXPERIMENT
TRAIN
TUNE
TRAINED MODEL
DEPLOYMENT STAGE
DEPLOY
PREDICTION/ INFERENCE
LOG
USE CASE DEPENDENT: FEATURE STORE
METADATA STORE / MODEL REPOSITORY
MONITORING (TRAINING - DEPLOYMENT)
USE CASE DEPENDENT: SYNTHETIC DATA GENERATION/ AUGMENTATION
AI/ML DATA FOUNDATION - VERSIONING DATA LAKE WITH LINEAGE TRACKING
AI Infrastructure Alliance
HTTPS://AI-INFRASTRUCTURE.ORG

Source: AI Infrastructure Alliance[110]

D. Applications

[109] David Gray Widder, Sarah West, & Meredith Whittaker, *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI* (draft), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807.

[110] ODSC Community, *The Rapid Evolution of the Canonical Stack for Machine Learning*, OPENDATASCIENCE, July 13, 2021, https://opendatascience.com/the-rapid-evolution-of-the-canonical-stack-for-machine-learning/. *See also* Giancarlo Mori, *Demystifying the Modern AI Stack*, MEDIUM, Nov. 1, 2022, https://gcmori.medium.com/demystifying-the-modern-ai-stack-d91ce73ec4e

Finally, we come to layer in the AI technology stack that is most visible to the public—the application layer. Consumers interact with AI through applications. For example, if I use ChatGPT to help me describe the application layer of the AI technology stack, I do so by entering a prompt into the ChatGPT application (say, "Describe the application layer of the AI technology stack in one sentence."). That application then interacts with a version (through prediction or inference) of the GPT model, and returns a result ("The application layer of the AI technology stack is responsible for providing end-users with specific AI-powered functionalities and services, utilizing various machine learning models and algorithms to solve particular tasks or address specific problems.").[111]

The industrial organization of the application layer includes three categories. The first are vertically-integrated applications. As in the ChatGPT example above, a single entity—OpenAI—has developed both the application and the underlying model. Similarly, Microsoft, which is a significant investor in OpenAI, has incorporated GPT into a wide range of its products, from Bing, its search engine, to Microsoft Office, among other products.[112] In some cases, these models are closed to third-parties because of the sensitive nature of the model and its underlying data. For example, the only available applications for certain AI-powered health applications are vertically integrated with the model itself.[113]

The second category are applications developed by unaffiliated third-party developers, who build upon existing proprietary foundation models. For example, some developers are using OpenAI's documented APIs to develop specific applications based upon GPT, such as patent drafting and analysis applications.[114] Notably, the firms who operate the foundation models could themselves set up applications that compete with third-party developers, and can set the terms on which data from the application is incorporated into future iterations of the underlying model. The third category are applications developed by third-party developers who rely on open source models and data. There are, for example, a range of developers using LLaMA's open source model (or other foundation models) to develop other language-based applications, including customer service chatbots.[115] In this category, the applications do not depend on vertically-integrated firms—except, perhaps, at the hardware and cloud layers.

Across all these organizational forms, we emphasize that inference (i.e., calling on a model to resolve a particular query) is typically relatively cheap—

[111] OpenAI, ChatGPT, at http://chat.openai.com?model=text-davinci-002-render-sha (prompt="Describe the application layer of the AI technology stack in one sentence.")

[112] Frederic Lardinois, *Microsoft Launches the New Bing, with ChatGPT Built In*, TechCrunch, Feb 7, 2023, at https://techcrunch.com/2023/02/07/microsoft-launches-the-new-bing-with-chatgpt-built-in/, Indeed, this integration extends beyond the model layer and into other layers, as OpenAI used Microsoft's cloud computing platform, Azure, to develop its GPT models. Microsoft, Official Microsoft Blog, Microsoft and OpenAI Extend Partnership, Jan. 23, 2023, at https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/.

[113] Arti K. Rai, Isha Sharma & Christina Silcox, *Accountability, Secrecy, and Innovation in AI-Enabled Clinical Decision Software*, 7 J.L. & BIOSCIENCES, Jan.–June 2020, at 1, 3, 5

[114] *See* Garden Intel., at https://www.gardenintel.com/ .

[115] *See* Ada, at https://www.ada.cx/.

especially when compared to the costs of model development and training.[116] But even if inference can seem relatively low-cost, inference at scale—resolving thousands or millions of queries—still requires a substantial resource investment.

II. The Drawbacks of An AI Oligopoly

Understanding the industrial organization of AI and the market structure of each layer in AI's technology stack shows that portions of the AI technology stack will be—and many already are—dominated by a small number of firms.[117] Concentration in the AI sector—an AI oligopoly—has a variety of downsides. In this Part, we outline four sets of problems with the AI oligopoly—abuses of power, national security and resilience issues, widening economic inequality, and effects on democracy.[118]

A. Abuses of Power and Economic Harms

As in other areas in which technology platforms dominate—operating systems, search, e-commerce, social media—concentration in AI seems likely to lead to a variety of abuses of power. Although widespread adoption of AI is only recent, any abuses of power by AI-related companies are likely to follow familiar and recognizable pathways.

1. Hardware — Given the structure of the microprocessor industry, customers may suffer from the problems of monopoly or oligopoly control. With only one firm in photolithography and a predominant firm in semiconductor manufacturing, it is quite possible that these firms could abuse their market power. They could demand monopoly prices for their goods, set discriminatory prices and terms for different customers, or refuse to deal entirely with some customers.[119] While it is not obvious that ASML has taken any of these actions so far, these strategies are often deployed by monopolists, to the detriment of the market.[120] With respect to TSMC, the semiconductor manufacturer has prioritized its contracts and partnership with Apple over other chip consumers, giving lower priority to service, networking, and PC chips during periods of shortages.[121] Indeed, Apple has reportedly

[116] *See* Narechania, supra note __, at 1580–81.

[117] Ajay Agarwal, Joshua Gans, & Avi Goldfarb, Prediction Machines: The Simple Economics of Artificial Intelligence 216 ("For technology companies whose entire business might rest on AI, scale economies might result in a few dominant companies.").

[118] We emphatically do not mean to say these are the only problems with AI or to offer a prioritization of all problems with AI.

[119] W. Kip Viscusi, Joseph E. Harrington, Jr. & John M. Vernon, Economics of Regulation and Antitrust 82 (4th ed. 2005)

[120] For a discussion see, Ricks et al., *supra* note xx, passim (discussing these abuses through a variety of industries, including railroads and operating systems).

[121] Samuel Nyberg, *Apple Gets Special Treatment Amid Chip Shortage*, MacWorld (Jun. 22, 2021, 1:01pm PDT), https://www.macworld.com/article/677141/apple-gets-special-treatment-amid-chip-shortage.html

"locked up" TSMC's entire capacity for fabricating 5 nanometer chips, which are currently the smallest and most advanced.[122]

The dynamics of GPUs are different. While Nvidia is far and away the dominant producer, the biggest technology companies—Amazon, Google, Meta, and Microsoft—are developing their own proprietary alternatives. Hence, these companies face some tension with Nvidia: they rely upon Nvidia's GPUs to run their supercomputers; but they are also simultaneously trying to reduce or even eliminate their reliance on Nvidia by developing independent alternatives. Indeed, if they can develop microchips that satisfy their training or inference requirements, they could, over time, vertically integrate further.

Here, it is worth distinguishing between training and inference. Given the significant processing power required for training new foundation models, it may be that even big technology companies will continue to rely on Nvidia for GPUs that are best suited to train models at the lowest cost. Inference, however, requires significantly less processing capacity—and can benefit from ASICs if the tasks are repetitive and predictable.[123] It is possible, perhaps even likely, that big technology companies will be able to integrate more deeply with respect to inference. In short, these providers may move down the stack, at least in part, in order to reduce their reliance on—vulnerability to—some of the dominant hardware providers.[124] And in response, it seems that chip designer Nvidia is also moving up the stack, offering their own cloud computing services.[125]

[122] Jeremy Horwitz, *Apple Blamed IBM and Intel for Mac Chip Delays, but TSMC Won't Be Next*, VENTURE BEAT (Nov. 13, 2020, 1:36 PM), https://venturebeat.com/mobile/apple-blamed-ibm-and-intel-for-mac-chip-delays-but-tsmc-wont-be-next/

[123] Andrej Karpathy, *Software 2.0*, MEDIUM, Nov. 11, 2017, https://karpathy.medium.com/software-2-0-a64152b37c35 (noting that one benefit of neural networks, which are used for AI, is that they can be programmed into a chip).

[124] Indeed, Google is making moves to design TPUs in-house, instead of relying on Broadcom. Wayne Ma, Anissa Gardizy & Jon Victor, *To Reduce AI Costs, Google Wants to Ditch Broadcom as its TPU Server Chip Supplier*, THE INFORMATION (Sept. 21, 2023, 1:22 AM PDT), https://www.theinformation.com/articles/to-reduce-ai-costs-google-wants-to-ditch-broadcom-as-its-tpu-server-chip-supplier.

[125] Nvidia, *NVIDIA Launches DGX Cloud, Giving Every Enterprise Instant Access to AI Supercomputer From a Browser*, Mar. 21, 2023, https://nvidianews.nvidia.com/news/nvidia-launches-dgx-cloud-giving-every-enterprise-instant-access-to-ai-supercomputer-from-a-browser. Perhaps because they do not have cloud infrastructure built up, Nvidia is or plans to partner with Oracle, Microsoft, and Google to provide this service. It is/has also bought Lambda Labs as a way to get into cloud provision directly. *See* Maria Heeter, Kate Clark & Stephanie Palazzolo, *Nvidia Accelerates AI Startup Investments, Nears Deal with Cloud Provider Lambda Labs*, THE INFORMATION (July 18, 2023, 5:00 AM PDT), https://www.theinformation.com/articles/nvidia-accelerates-ai-startup-investments-nears-deal-with-cloud-provider-lambda-labs; Anissa Gardizy, *In an Unsual Move, Nvidia Wants to Know Its Customers' Customers*, THE INFORMATION (July 31, 2023, 6:00AM PDT), https://www.theinformation.com/articles/in-an-unusual-move-nvidia-wants-to-know-its-customers-customers. Nvidia is also leasing servers powered by its own chips in Google Cloud Platform and among other cloud providers, a development that has been called a "trojan horse" and an effort to "muscle" its way into the lucrative business. Anissa Gardizy & Aaron Holmes, *Nvidia Muscles into Cloud Services, Rankling AWS*, THE INFORMATION (Sept. 11, 2023, 1:17 PDT), https://www.theinformation.com/articles/nvidia-muscles-into-cloud-services-rankling-aws.

These developments invite an evaluation of three different possible market structures (assuming, for now, that higher layers in the stack —i.e., cloud computing, AI models, and applications—are competitive). The first is a monopoly market structure, characterized by one predominant GPU provider. The second is a competitive market structure, in which multiple providers each sell to distinct corporate entities in—i.e., structurally separated from—higher layers in the stack. That is, GPU production is competitive and GPU producers do not also own, operate, or hold investments in higher layers in the stack. The third is a vertically-integrated market structure, in which there are multiple GPU producers who are vertically integrated with cloud provision and other layers in the stack. Until now, a monopoly structure has characterized the hardware layer, with Nvidia the predominant provider of processing hardware. But the developments we describe above are suggestive of at least tentative shifts towards a vertically-integrated market structure.

Both the monopoly and vertically-integrated market structures present competition concerns. The monopoly structure because the GPU producer is a monopolist or has such significant market power that it could raise consumer costs, price certain users out of the market, discriminate against certain purchasers, or impede new competition. Indeed, there are reports that Nvidia's current chip allocation decisions are based on whether it is "excited about [the] end customer" in part because "Nvidia would prefer not to give large allocations to companies that are attempting to compete directly with them."[126] The vertically-integrated structure poses risks to competition because vertical integration could entrench the existing oligopoly, making it harder for new competitors to emerge. New competitors would face significant barriers to entry—high capital costs of chip acquisition (not to mention the even higher costs of design and production). Moreover, the existing players have a strong incentive to lock out putative competitors.[127] And because hardware and cloud infrastructure would be integrated, new entrants would have to operate in both layers in order to compete effectively. Entities in the higher layers—model or application developers—may not otherwise be able to gain access to necessary infrastructure if the entrenched oligopoly requires that these users purchase an integrated (i.e., hardware *and* cloud infrastructure) service.[128] The

[126] Clay Pascal, *Nvidia H100 GPUs: Supply and Demand*, GPU UTILS, July 2023, Updated August 2023, https://gpus.llm-utils.org/nvidia-h100-gpus-supply-and-demand-.

[127] Even skeptics of monopoly leveraging theories (due, for example, to the one-monopoly-profit theory) might be persuaded by the possibility for leveraging in this context. JONATHAN E. NUECHTERLEIN & PHILIP J. WEISER, DIGITAL CROSSROADS: TELECOMMUNICATIONS LAW AND POLICY IN THE INTERNET AGE (2d ed. 2013) 14–17 (describing the theory and exceptions to it); Philip J. Weiser, *Toward a Next Generation Regulatory Strategy*, 35 LOY.-CHI. L.J. 41, 73 (2003) ("[T]here are instances in which a platform provider may use its gatekeeping role to 'hold up' the deployment of applications, thereby giving itself an additional source of revenue and deterring future innovation.")

[128] *Cf.* Reg. & Policy Problems Presented by the Interdependence of Computer and Commc'n Servs. & Facilities, Tentative Decision of the Commission, 28 F.C.C. 2d 291, para. 24 (1970) (discussing a similar concern in the related computer networking and information processing context); Lina M. Khan, *The Separation of Platforms and Commerce*, 119 COLUM. L. REV. 973 (2019) (discussing a similar concern in the related platforms context).

competitive structure, in contrast, offers a robust competitive environment between the two layers.

Nevertheless, there may be reasons to favor a vertically-integrated market. Microprocessing is, after all, tightly connected to the rest of the computing hardware (including the hardware used to deliver cloud services), and so vertical integration may yield substantial benefits. Nvidia runs supercomputers, in part, because talented engineers want to be able to work on the supercomputers, not just design GPUs.[129] As importantly, the fact that ASICs can be developed for specific inferential tasks suggests advantages for integrating of these hardware layers with model and application layers.

This question—whether the technological connections are so tightly linked such that vertical integration is preferable—echoes in the early debates of network neutrality and, especially, open access to the cable industry's broadband networks. There, some advocates argued that the cable industry's networks should be made open to competing ISPs, such that not only Comcast—but also America Online and RoadRunner, among others—could all offer service over a single set of wires. Yet others countered that offering effective broadband internet service required control over the infrastructure, as such control enabled ISPs to configure the hardware to improve performance. Hence, the debate now—much as it was then—regards whether vertical integration or greater competition is, on net, better for downstream applications.[130] The answer to this question remains uncertain, and it is somewhat hard to disentangle the providers' technical arguments favoring integration from their economic incentives: They might easily deploy the former (accurately or not) in service of the latter. But an appropriate regulator, aided with relevant expertise, empowered to collect technical information, and authorized to address the concerns of concentration and integration, might be able craft an appropriate response, drawing from the proposals we set out in Part IV.

2. *Cloud Infrastructure* — The concentration of cloud providers has significant consequences for competition—and for the future of AI. The dominant cloud providers have taken steps to entrench their dominance, including by facilitating lock-in effects that raise the costs for consumers to switch providers.[131] In addition to the lack of interoperability and the need for expertise in each system, multi-year contracts and egress fees all impede competition in the market.[132] Providers have also vertically integrated across higher and lower layers of the technology stack, enabling them to offer more—and more tailored—services.

While more integrated offerings might seem beneficial, they also come with downsides for users, third-party developers, and society at large. For users, concentration and lock-in means high prices. Although cloud computing is, at bottom, largely a commodity product—computational capacity's cost has decreased

[129] Nicole Kobie, *Nvidia and the Battle for the Future of AI Chips*, WIRED, June 17, 2021 (10:10AM), https://www.wired.co.uk/article/nvidia-ai-chips.

[130] Tim Wu, *Network Neutrality, Broadband Discrimination*, 2 J. ON TELECOMM. & HIGH TECH. L. 141 (2003).

[131] *See* Federal Trade Commission, Solicitation for Public Comments on the Business Practices of Cloud Computing Providers, No. 2023-0028-0001 (March 22, 2023).

[132] Investigation into Digital Markets, *supra* note xx, at 98-99.

over time[133]—cloud providers like AWS can charge substantial premiums (e.g., 30 percent margins) on this service.[134] Andreesen Horowitz, one of the most notable technology investment firms, argues that these cloud fees are so substantial that many companies would be better off providing these services in-house—that is, many companies would save money by building their own internal cloud platform. Andreesen Horowitz estimates that the top 50 public software companies could recover about $100 billion in market capitalization from the cost differential between providing in-house cloud and using one of the big vendors.[135] This is due to the "cloud paradox"—start-up companies must employ external cloud vendors because of the high capital costs of developing the service; once established, such companies should prefer proprietary service over these higher cost external vendors—but they stick with the higher cost approach.

For third party firms, reliance on cloud services can mean vulnerability to copying and self-preferencing by the cloud provider. Multiple firms have complained that AWS has copied their product and offered their own integrated version of the product, harming their company's value and future business.[136] The prospect of expropriation of creativity and effort by a cloud provider may not only lead entrepreneurs and venture funders to prefer not to invest in innovative companies, it may deter such innovative activity altogether—particularly if such conduct is pervasive across a concentrated set of dominant service providers. After all, why would anyone invest in a new venture, when the dominant cloud provider is likely to just copy the idea and integrate it into their platform?[137] Even if the platform does not copy the firm's business but instead acquires it early on, this may also reduce incentives for venture funders, as they do not get the financial upside of investing in a more successful company. Venture capitalists call this the "kill zone," and leading economists have modeled its existence in internet platform markets.[138]

Dominant cloud platforms can also leverage their power from cloud services into other parts of the AI stack. Looking to the lower levels of the stack, Amazon, Microsoft, and Google are, as noted, developing microprocessing units specific to AI applications in order to fully integrate their hardware components

[133] Paddock, *supra* note xx.

[134] Sarah Wang & Martin Casado, *The Cost of Cloud, a Trillion Dollar Paradox*, ANDREESSEN-HOROWITZ, https://a16z.com/2021/05/27/cost-of-cloud-paradox-market-cap-cloud-lifecycle-scale-growth-repatriation-optimization/ (last visited July 11, 2023).

[135] *Id*.

[136] Andrew Leonard, *Amazon Has Gone From Neutral Platform to Cutthroat Competitor, Say Open Source Developers*, ONEZERO (April 24, 2019) https://onezero.medium.com/open-source-betrayed-industry-leaders-accuse-amazon-of-playing-a-rigged-game-with-aws-67177bc748b7; Jordan Novet, *Amazon Steps Up its Open-Source Game, and Elastic Stock Falls as a Result*, CNBC (March 12, 2019) https://www.cnbc.com/2019/03/12/aws-open-source-move-sends-elastic-stock-down.html; Jordan Novet, *Amazon's Cloud Business is Competing with its Customers*, CNBC (Nov. 30, 2018), https://www.cnbc.com/2018/11/30/aws-is-competing-with-its-customers.html. This provides two additional examples of Amazon introducing services that copy and compete with companies reliant on their cloud infrastructure.

[137] *See* RICKS ET AL, *supra* note xx, at 15-16 for a discussion of this point.

[138] Sai Krishna Kamepalli, Raghuram Rajan & Luigi Zingales, *Kill Zone*, NBER Working Paper, https://www.nber.org/papers/w27146.

across chips and cloud. Looking to the higher levels, they also all offer applications that could integrate AI models. Microsoft's Bing search engine and Office 360 suite, for example, can integrate AI inference, built atop OpenAI models (OpenAI is funded, in significant part, by Microsoft), and run on Microsoft Azure compute power. Google and Amazon can do the same with their various offerings, from search to e-commerce. Integration across these domains makes it harder for new entrants to compete at these other stages. Can a new search engine compete without having its own AI model to improve search capacity? Could a word processor compete with Microsoft Office without its own, integrated ChatGPT-type system? As consumers come to expect these features, effective competition will require that putative competitors develop offerings across the entire stack, thereby increasing barriers to entry at nearly any layer. Indeed, when rivals in search who license Microsoft's system have tried to use it for training their own AI models, Microsoft has threatened to block their access to the data, as a violation of its terms of service.[139]

3. *The Model Layer* — The technical and market characteristics of the model layer highlight several possible problems with concentration in the AI sector. We can begin with the data that underlies a model's development. As noted, while some data is freely and easily available, it may require significant resources to transform and label. And while there may be some vast troves of data, it can be expensive to obtain *good* data—data that is, say, debiased in ways that are critical to the development of fair and accurate downstream applications. Other data is proprietary and expensive, presenting a significant barrier to entry. And new entrants face a growing challenge: As AI systems rocketed in popularity, previously free sources of training data have since limited access to this information, or are now seeking to monetize it for AI training purposes.[140]

Moreover, there may be significant data network effects for some models, particularly for models and applications that rely on forms of deep, continual, or reinforcement learning, giving rise to significant first-mover advantages. As one scholar has written elsewhere, models

> that continue to internalize new data, including information drawn from their practical deployments, may gain an insurmountable lead over putative competitors in their initial competition for the market. This is because the first application in the market gains access to more recent and more relevant training data—data from in market consumers—before any competitor. Integrating those results into its prediction scheme thus gives rise to better results for the next query. And that next query, again, gives the provider

[139] REUTERS, *Microsoft Threatens To Restrict Data From Rival AI Search Tools*, March 25, 2023, https://www.reuters.com/technology/microsoft-threatens-restrict-data-rival-ai-search-tools-bloomberg-news-2023-03-25.

[140] *See* Mike Isaac, *Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems*, N.Y. TIMES, April, 18, 2023, at https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html.

> even more recent and relevant data that may further improve its application—and so on.[141]

Leading members of the industry have likewise observed that this process is a "virtuous cycle for strengthening the best products and companies" and that AI thus appears as a "winner-take-all" system.[142] In short, scale can matter a lot to data,[143] and scale is becoming harder to achieve.[144]

The barriers to entry for foundation model development extend beyond data to include, as noted, the significant compute resources that are required to train a model. Taken together, these barriers suggest that, in some—perhaps many—fields, only one or few foundation models are likely to emerge per application class.

This concentrated control over foundation models gives rise to both competition and quality concerns. On quality, as the concentration among publicly available foundation models increases, the quality of each one matters ever more. If there are only one or two models in an application class (e.g. text, images) and they are flawed, then every application built on that model will suffer from those flaws.[145] The stakes are significant. With respect to competition, the monopoly or oligopoly structure for foundation models—which, as their name suggests, are foundational to further AI development—gives rise to familiar concerns regarding concentration in other platform markets.[146] Foundation model providers might raise the costs to downstream developers for model access. They may favor selected AI-based applications, including vertically integrated applications, through selectively exposed APIs.[147] And they might even copy applications from competitors and incorporate them into their own offerings.

[141] Narechania, *supra* note xx, at 1584.

[142] KAI-FU LEE, AI SUPERPOWERS: CHINA, SILICON VALLEY, AND THE NEW WORLD ORDER 19-20 (Mariner Books ed., 2021) (2018).

[143] AJAY AGARWAL, JOSHUA GANS, & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE 49–50 (citing arguments that "[i]ncreasing data brings disproportionate rewards in the market"); see also *id*. at 216 (describing scale advantages for long-tail instances).

[144] *See* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM 187 (2019) ("[M]achine learning is only as intelligent as the amount of data it has to train on, and Google has the most data.").

[145] For a notable discussion on how the size of LLMs does not lead to diversity of outputs, see Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Schmargaret Schmitchell, *On the Danger of Stochastic Parrots: Can Language Models be Too Big?*, FAccT '21, March 3-10, 2021, Virtual Event, Canada. https://dl.acm.org/doi/pdf/10.1145/3442188.3445922.

[146] For a discussion of Foundation Models, including of competition concerns, see U.K. COMPETITION AND MARKETS AUTHORITY, AI FOUNDATION MODELS: INITIAL REPORT 27-53 (Sept. 18, 2023), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1185508/Full_report_.pdf.

[147] Such concerns have long been at the heart of related debates, including, as noted, debates over network neutrality. *See, e.g.*, Protecting and Promoting the Open Internet, 30 FCC Rcd. 5601, 5632–33 (2015). *See also* Philip J. Weiser, *The Internet, Innovation, and Intellectual Property Policy*, 103 COLUM. L. REV. 534, 579 (2003) ("In the government's antitrust case against Microsoft, for example, the government submitted evidence of a manager's statement that 'to control the APIs is to control the industry' and established that Microsoft's monopoly rested, in part, on its firm control of its APIs.'"); *cf. supra* note 128

4. Applications — At the application layer, potential abuses of power also follow from vertical integration, as they do in the model layer. As we have seen, some companies have vertically integrated across the entire AI technology stack. Where these companies have exclusive access to proprietary data (e.g. medical information), vertical integration will likely limit competition downstream in the model or application layers. Where these companies offer their APIs to developers to create third party applications (as, for example, OpenAI does at present), the model provider might also create its own competing applications. This raises a variety of anti-competitive concerns.[148] They might exclude some third-party applications from use of the model. Model providers might favor their own applications over others by charging higher rates to third-party developers than their own in-house business lines, or by conferring other advantages on their own products.[149] For example, if people ask Microsoft Bing what they should do this weekend, it might suggest playing video game Call of Duty—which Microsoft also owns.[150] Or they might copy third-party applications and develop and integrate those features into their own applications. And while such integration may not seem problematic at first blush, we note two concerns with such conduct. First, such integration immediately raises the concerns about foreclosure and favoritism that we have just identified. Second, the possibility that an integrated entity will copy and integrate the features of a popular application diminishes the likelihood that anyone will invest in application creation at all, leading to fewer new applications overall. Such anticompetitive conduct has lengthy pedigree across a range of NPU sectors—including technology platforms.[151]

As we have seen, the downstream effects of these abuses of power are potentially significant. Applications developers may decline to develop new applications based on foundation models, if they believe the model providers will steal and copy their idea. Venture firms increasingly see the layer as a "kill zone," subject to lower returns on investment (as compared to a more competitive ecosystem across the stack), thereby reducing overall investment in the sector. And a lack of competition in the application layer will reduce innovation.

B. National Security and Resilience

Concentration at critical points in the AI technology stack also raises significant concerns from a national security and resilience perspective. Consider the supply of microprocessing units. With very few chip companies—and particularly semiconductor and photolithography manufacturers—the possibility that one

[148] *See* Tom Slee, *The Incompatible Incentives of Private-Sector AI, in* THE OXFORD HANDBOOK OF ETHICS OF AI 107, 122 (Markus D. Dubber, Frank Pasquale & Sunit Das eds., 2020) ("Algorithmic ranking systems can beomce power institutions in and of themselves: part of the infrastructure of society. Advantages accrue to the company that owns the infrastructure when it is also competing the market for services that exploit that infrastructure.").

[149] *See supra* notes 128, 147 and accompanying text.

[150] We are indebted to Nick Garcia of Public Knowledge for this example.

[151] *See, e.g.*, RICKS ET AL., 475-532 (discussing railroads), 935-970 (discussing computer operating systems).

foundry could be shut down due to a pandemic, weather event, war, or other emergency is—and, indeed, has been—significant.[152] Concentration leads to a fragile supply chain that is vulnerable to single points of failure.[153] More specifically, there are clear national security concerns with respect to the supply chain for chips.[154] Given that chips power not only AI but other critical technologies, the lack of availability could impede both military and non-military critical infrastructure.[155] TSMC's dominance in manufacturing has led to concerns about what might happen if China attempts to take over Taiwan or if the U.S. and China get into a conflict.[156]

In addition to resiliency concerns, those focused on global competition and international leadership have observed that staying ahead on technology will be critical to power in the 21st century.[157] In this context, the dominance of a single company in semiconductor manufacturing—and a company located in Taiwan—raises risks. A more diverse supply chain—both geographically and among multiple firms, and including U.S. production—would help ensure American global leadership in cutting-edge technology.

Other layers in the AI technology stack also raise national security and resilience issues. An oligopoly of cloud providers, integrated up and down the AI stack and without interoperability between them, gives rise to substantial software supply-chain concerns.[158] If a cloud provider is attacked in a cyberattack, or if a cloud provider's warehouse is affected by a severe weather event, or even if an employee makes a simple mistake, dozens of AI applications—and the operations, services, and websites that depend on them—could shut down for hours, days, or longer.[159] Such disruptions would not only harm the affected companies but could have devastating effects on the economy as a whole.[160] The lack of interoperability means that these systems could not easily be restarted on another provider's

[152] *See, e.g.*, Karen M. Sutter, John F. Sargent, Jr. & Manpreet Singh, *Semiconductors and the CHIPS Act: The Global Context* 2, CONG. RES. SERV., May 18, 2023, https://crsreports.congress.gov/product/pdf/R/R47558

[153] On economic resilience as a critical part of foreign policy, see Ganesh Sitaraman, *A Grand Strategy of Resilience*, FOREIGN AFFAIRS (Sept./Oct. 2020).

[154] For a history of the relationship between foreign policy and chips, see CHRIS MILLER, CHIP WAR (2022).

[155] ID. at 327-334.

[156] ID. at 335-344.

[157] Eric Schmidt, *Innovation Power*, FOREIGN AFFAIRS (Mar./Apr. 2023).

[158] *See* U.S. Treasury Dept., *New Treasury Report Assesses Opportunities, Challenges Facing Financial Sector Cloud-Based Technology Adoption* (Feb. 8, 2023), https://home.treasury.gov/news/press-releases/jy1252 ("The current market is concentrated around a small number of CSPs [Cloud Service Providers], which means that if an incident occurs at one CSP, it could affect many financial sector clients concurrently.").

[159] *See, e.g.*, Nick Merrill & Tejas N. Narechania, *Inside the Internet*, DUKE L.J. (forthcoming 2023).

[160] U.S. Treasury Dept., *New Treasury Report Assesses Opportunities, Challenges Facing Financial Sector Cloud-Based Technology Adoption* (Feb. 8, 2023), https://home.treasury.gov/news/press-releases/jy1252 ("Many financial institutions have expressed concern that a cyber vulnerability or incident at one CSP [Cloud Service Provider] may potentially have a cascading impact across the broader financial sector.").

service. Hence, for the U.S. government and military, these owners of cloud computing infrastructure are mission critical providers of national infrastructure.

Concentration at the foundation model layer can also lead to national security concerns. Imagine, for example, a single foundation model for certain medical diagnoses, in which the data or training system is flawed and leads to plausible but incorrect outputs. It is possible that widespread use of such a model could systematically lead to misdiagnoses and mis-prescribing remedies. Perhaps, during normal time, regulatory processes and protections, or competition, would suffice to catch these errors. But if only one firm has the capacity to deploy such technologies in an emergency—during, say, a pandemic—an error in this concentrated ecosystem could be catastrophic. For the military, reliance on a single foundation model for any number of activities—from design of military hardware, to automated responses—could have unintended and deadly effects. Concentration in the AI technology stack makes this phenomenon worse: there may be a severely limited number of providers, and therefore little ability to switch toward one with better service.

C. Economic Inequality

Concentration at layers within and across the AI technology stack can also deepen economic inequality in at least three ways. First, concentration means that a small number of firms will capture the vast majority of the financial returns in this sector. As technologist and investor Kai-Fu Lee puts it, "[c]orporate profits will explode, showering wealth on the elite executives and engineers lucky enough to get in on the action…."[161] For the United States, which is already on the high end of historic economic inequality in the population,[162] continuing the concentration of income and wealth both arrests economic mobility[163] and is undesirable for those who seek a more egalitarian society.

Second, as AI is deployed, it could create a "bifurcated job market that squeezes out the middle class."[164] While it is too early to tell exactly how the labor market will change, it is likely that jobs that are not easily routinized and require in-person, physical, customized interactions, such as child care or elder care, are more likely to remain insulated from technological replacement or augmentation. Other jobs, however, that rely on repetitive, routinized drafting or other tasks for which AI is suited could require less labor over time.[165] Some jobs will become

[161] LEE, *supra* note xx, at 171; *see also* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM 500 (2019) (noting that "GM employed more people during the height of the Great Depression than either Google or Facebook employ[ed] at their heights of market capitalization.").

[162] *See, e.g.*, THOMAS PIKETTY, CAPITAL IN THE TWENTY-FIRST CENTURY (2013); Emmanuel Saez and Gabriel Zucman, *The Rise of Income and Wealth Inequality in America: Evidence from Distributional Macroeconomic Accounts*, 24 J. ECON. PERSP. 3 (Fall 2020).

[163] For an overview of this argument and data, see Jared Bernstein & Ben Spielberg, *Inequality Matters*, THE ATLANTIC, June 5, 2015, https://www.theatlantic.com/business/archive/2015/06/what-matters-inequality-or-opportuniy/393272/.

[164] LEE, *supra* note xx, at 170

[165] Acemoglu & Restrepo, *supra* note xx.

human-plus-AI jobs,[166] increasing the productivity of individual workers, but requiring fewer workers overall. It is possible that, over the long run, shifts in the type of labor—like the shift from agriculture to factories in the early 20th century—will work themselves out. But in the short to medium term, there can be extraordinary pain for workers who lose their jobs and struggle to find new ones.[167] Research studying the people and areas that faced substantial job losses from trade and offshored production in the 2000s show that those areas did not bounce back—even after years.[168] And these changes affect the autonomy of those workers who are the object, rather than the subject, of this change.[169] These consequences could also have second-order effects: reshaping people's views and mobilizing individuals toward destabilizing political change.[170]

Third, concentration in AI is likely to increase global inequality, as the dominant firms, located in a small number of industrialized and technologized countries, extract value from data that is harvested from other economies.[171] For those who are concerned about the economic well-being of peoples and nations around the world, the concentration of economic benefits within a small number of countries is a problem. And, looking beyond economic considerations, the divide in AI development across the so-called "Global North" and "Global South" may have important cultural implications, as predominantly English-based systems accelerate the threats, for example, to endangered languages.[172] In all, as Lee concludes, not only will "AI-rich countries . . . amass great wealth," but those countries will also "witness the widespread monopolization of the economy and a labor market divided into economic castes."[173]

[166] For a discussion of how regulation can lead to harnessing technologies in this manner, see FRANK PASQUALE, NEW RULES OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI (2020).

[167] Acemoglu & Restrepo, *supra* note xx; Anton Korinek, *Integrating Ethical Values and Economic Value*, in THE OXFORD HANDBOOK OF ETHICS OF AI 475, 480–83 (2020).

[168] David H. Autor, David Dorn & Gordon H. Hanson, *The China Shock: Learning from Labor Market Adjustment to Large Changes in Trade*, 8 ANN. REV. ECON. 205 (2016).

[169] Korinek, supra note __, at 486–487; cf. Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as Object*, 52 STAN. L. REV. 1373, 1426 (2000).

[170] See STEWART RUSSELL, HUMAN COMPATIBLE 113–124 (2019) (describing the "risk [of] an unsustainable level of socioeconomic dislocation").

[171] Steven Weber & Gabriel Nicholas, *Data, Rivalry and Government Power: Machine Learning Is Changing Everything*, 14 GLOB. ASIA 23 (2019)

[172] Victoria Marian, *AI Could Cause a Mass-Extinction of Languages — And Ways of Thinking*, WASH. POST, April 19, 2023, at https://www.washingtonpost.com/opinions/2023/04/19/ai-chatgpt-language-extinction/. For a more general argument, see generally Fleur Johns, *Data Mining as Global Governance, in* THE OXFORD HANDBOOK OF LAW, REGULATION, AND TECHNOLOGY 776 (Roger Brownsword, Eloise Scotford & Karen Yeung eds., 2017).

[173] LEE, *supra* note xx, at 172.

D. Democracy

The concentration of economic power has long been understood as a danger to a republican form of government.[174] In the AI context, concentration in and across the technology stack raises concerns for the health of our democracy.[175] For starters, democracy depends on vibrant political debate and discussion.[176] Concentration in the number of foundation models—and in vertically-integrated applications—can shape the information ecosystem in profound ways, emphasizing certain topics of conversation. Indeed, concentration in the AI stack is not independent from algorithmic decision-making. If there are only a few information sources that rely on a small number of foundation models, model providers are likely to have an outsized influence on information. Private and individual control may both be problematic. The former because private firms are guided by private interests, rather than the public good, and so may have an interest in facilitating information that is financially beneficial even if otherwise problematic. The latter because individuals might have ideological or idiosyncratic aims. In either case, an AI oligopoly concentrates power in a way that could be dangerous to a diverse speech ecosystem and thus to democratic government. Moreover, AI may reduce the cost for malicious actors wishing to shape the information ecosystem to their particular ends, such as by facilitating and amplifying the distribution of "deepfakes" (i.e., AI-generated photos, images, or videos that seem to reflect actual (but in fact falsified) events).[177]

Economic power also often translates into political power. Corporate lobbying shapes the political system in a range of ways: from agenda control to substantively forestalling regulation.[178] Importantly, corporate lobbying power does not just apply to the sector in which the companies operate. Powerful companies also lobby about general economic policies—from tax and labor issues to regulatory issues outside their domain.[179] Importantly, political scientists have shown that

[174] *See generally* GANESH SITARAMAN, THE CRISIS OF THE MIDDLE-CLASS CONSTITUTION: HOW ECONOMIC INEQUALITY THREATENS OUR REPUBLIC (2017) (describing the intellectual history of this point). For an overview of how economic power influences political and constitutional design—and is difficult to address, see Ganesh Sitaraman, *The Puzzling Absence of Economic Power in Constitutional Theory*, 101 CORNELL L. REV. 1445 (2016).

[175] *See* Sonia K. Katyal, *Democracy and Distrust in an Era of Artificial Intelligence*, DAEDALUS (THE J. OF THE AM. ACAD. OF ARTS & SCI.) (2022).

[176] See Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting); Zechariah Chafee, Jr., Freedom of Speech in War Time, 32 HARV. L. REV. 956–58 (1919); *see also* James Weinstein, *Participatory Democracy as the Central Value of American Free Speech*, 97 VA. L. REV. 491 (2011).

[177] Sayash Kapur & Arvind Narayanan, *Is AI Generated Disinformation A Threat to Democracy?*, AI SNAKE OIL, June 19, 2023; Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 127 CALIF. L. REV. 1753 (2019)

[178] *See, e.g.*, LEE DRUTMAN, BUSINESS OF AMERICA IS LOBBYING (2015).

[179] ALYSSA KATZ, THE INFLUENCE MACHINE: THE U.S. CHAMBER OF COMMERCE AND THE CORPORATE CAPTURE OF AMERICAN LIFE (2015).

companies and trade associations have an outsized influence on politics.[180] These are concerns across areas of policy, including in AI.[181]

As noted above, economic power at the individual level is also a form of political power. A voluminous literature in political science shows that wealthy individuals influence politics to a greater degree than those who are less wealthy. They participate at every stage of the political process to a greater degree.[182] When their preferences diverge from the majority's view, political scientists have shown that the wealthy's views usually hold: majority preferences have essentially no effect on policy outcomes.[183] To the extent that an AI oligopoly facilitates individual economic inequality it will also have effects on shaping political inequality and influence.

*

The drawbacks of an AI oligopoly—one that flows from AI's technical, industrial, and market organization—are substantial, implicating economic, national security, social, and political concerns. Concentration among service providers in the AI technology stack gives rise to concerns about price, quality and self-preferencing and discrimination, as well as questions about dynamic innovation. Such concentration also implicates resilience and security concerns, as these bottlenecks become critical single points of failure in our national security and economic infrastructures. And concentration also exacerbates concerns about economic inequality, and even for the future of democracy.

III. LESSONS FOR GOVERNANCE

Our account of the industrial organization of AI and the drawbacks of an AI oligopoly yields four important lessons. First, that the potential harms of an AI oligopoly are stable and independent of AI's ongoing development, and so there is little reason to wait before regulating. Second, that ex ante regulation ought to be seen as an essential mode of governance for this sector (as opposed to relying only on ex post enforcement). Third, the current trajectory of a vertically integrated AI oligopoly is likely to hinder innovation, and regulation can facilitate downstream innovation. And fourth, attention to AI's market structure is important for addressing the range of AI's potential harms—bias, false or misleading determinations, and so on—that have so far captured the attention of advocates, policymakers, and the public both because it focuses attention on where to regulate and because market concentration contributes to these harms.

[180] KAY LEHMAN SCHLOZMAN, SIDNEY VERBA & HENRY E. BRADY, THE UNHEAVENLY CHORUS 404-11 2012).

[181] *See, e.g.*, Yochai Benkler, *Don't Let Industry Write the Rules for AI*, NATURE, May 1, 2019, https://www.nature.com/articles/d41586-019-01413-1.

[182] SCHLOZMAN ET AL., *supra* note xx, at 13-21, 117-33.

[183] MARTIN GILENS, AFFLUENCE AND INFLUENCE: ECONOMIC INEQUALITY AND POLITICAL POWER IN AMERICA 97–123 (2012); LARRY M. BARTELS, UNEQUAL DEMOCRACY 253-54 (2008); Martin Gilens & Benjamin I. Page, *Testing Theories of American Politics: Elites, Interest Groups, and Average Citizens*, 12 PERSP. ON POL. 564, 573 (2014).

A. The Folly of Waiting to Solve Technology's Problems

Our analysis of the AI technology stack suggests that the problems we describe are a function of relatively stable, intrinsic characteristics. Stated otherwise, while the technology is developing rapidly, the industrial organization and concomitant market structure that flows from this technology is both easily discerned and a function of traits inherent to the technology and its industrial environment. And the pace of technological development does not affect these fundamentals or the harms that flow from them.

This finding has substantial implications for AI governance. One common response to proposals to govern AI (or any new technology, for that matter) is that the technology is too new, or is moving too quickly, for effective governance. As one analyst describes the objection, "Dealing with the velocity of AI-driven change . . . can outstrip the federal government's existing expertise and authority."[184] This is sometimes referred to as the "pacing problem," the idea that the pace of innovation is beyond the capacity of regulators.[185]

We disagree. To be sure, we do not mean to suggest that there are no outstanding questions. As we note above, one open question regards the benefits of integration across the hardware and cloud computing layers. But, in areas like this, public governance can sensibly account for the possibility that integration might be beneficial by, for example, declining to impose a separations rule between those layers at this time—while still protecting against other pitfalls of concentration in these layers through interoperability rules, cooperative structures, or the development of public cloud computing options.

Moreover, declining to regulate in view of ongoing technological development threatens to forestall regulation altogether. This is due to the so-called "Collingridge Dilemma:" If regulation is deemed unadvisable at the early stage of a technology because information is limited, once the technology becomes familiar, regulation becomes practically impossible because its proponents are entrenched.[186] In other words, the failure to regulate at the incipiency of a new technology means having to regulate after the industry has developed, when it has more political power to delay, weaken, or block any proposed regulation. As former FCC-chair Tom Wheeler has observed, taking a "self-regulatory approach" because of fears that government cannot regulate new technologies is tantamount to no regulation—and that is precisely what happened in online markets.[187] "The results of this strategy," he concludes, "speak for themselves in well-known current

[184] Tom Wheeler, *The Three Challenges of AI Regulation*, BROOKINGS, June 15, 2023, https://www.brookings.edu/articles/the-three-challenges-of-ai-regulation/

[185] Michael Guihot, Anne F. Matthew & Nicolas P. Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 VAND. J. ENT. & TECH. L. 385, 389 (2017) (discussing the pacing problem in the AI context); Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633, 653 (2020) (arguing against an FDA model for regulating AI because of the pacing problem, among other reasons). For a thoughtful discussion of regulating AI under conditions of uncertainty, see Maria Nordstrom, *AI under Great Uncertainty: Implications and Decision Strategies for Public Policy*, 37 AI & SOCIETY 1703 (2022).

[186] DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY 19 (1980). For an application to AI, see Guihot et al., *supra* note xx, at 422.

[187] *See* Wheeler, *supra* note xx.

online harms," including not only "market concentration," but also "invasion[s] of personal privacy," "user manipulation, and the dissemination of hate, lies, and misinformation."[188]

B. The Advantages of Ex Ante Governance

Our analysis of the AI technology stack and its market structure—coupled with an assessment of current law—also suggests that ex ante governance solutions will be superior tools for regulating AI than either a wait-and-see approach or ex post antitrust enforcement on a case-by-case basis.

Some have argued that we should embrace an approach of "permissionless innovation,"[189] allowing these companies to run amok until they cause substantial harm—and only then seek to redress it through forms of ex post enforcement, as under antitrust law. We disagree. Antitrust enforcement can be a powerful antimonopoly tool to address specific problems with abuses of market power and to shape markets and create deterrence. Indeed, as Tim Wu has argued, some of the biggest antitrust cases, even as they looked backwards at harms that had taken place, helped shape competitive markets by deterring anticompetitive behavior.[190] At the same time, antitrust law as currently interpreted and implemented is not likely to be sufficient. For those concerned about competition, innovation, and the harms of monopoly and oligopoly, ex ante regulatory tools will also be essential.

To see why, it is first important to understand how antitrust doctrines have been narrowly drawn, in ways that are likely to make it difficult for plaintiffs in the AI sector to win cases. Consider, for example, predatory pricing. Predatory pricing occurs when a firm sells its goods or services below cost in order to drive a competitor out of the market. Once the competitor has departed, the firm can then raise prices to supracompetitive levels. NPU sectors "may be particularly susceptible to predatory pricing" because of the winner-take-all dynamics of the businesses.[191] In the AI context, entrenched cloud providers might undercut new entrants with lower fees. Foundation model providers might do the same. Winning the market in these layers may be particularly valuable for firms because users of the platforms will face high costs for switching. The challenge, however, is that the Supreme Court has made it difficult for plaintiffs to win predatory pricing cases. The Court has been skeptical that predatory pricing ever takes place,[192] and has required plaintiffs to show that the defendant could likely recoup its losses—even in a case with clear evidence of predatory pricing.[193] This judicial skepticism may make predation cases less likely to be successful.

[188] *Id*.

[189] *See, e.g*., Adam D. Thierer, *Getting AI Innovation Culture Right*, April 13, 2023, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4404402 (arguing for a permissionless innovation approach).

[190] Tim Wu, *The Google Trial is Going to Rewrite Our Future*, N.Y. TIMES (Sept. 18, 2023), https://www.nytimes.com/2023/09/18/opinion/contributors/google-antitrust-trial.html.

[191] RICKS ET AL., *supra* note xx, at 226.

[192] Matsushita Electrical Industrial Co. v. Zenith Radio Group, 475 U.S. 574, 589 (1986)

[193] Brooke Group Ltd. v. Brown & Williamson Corp., 509 U.S. 209 (1993).

Another pricing doctrine focuses on "price squeezes," which occur when a vertically integrated firm with market power in an upstream business lines charges high prices to downstream competitors. Power in the upstream market allows the firm to benefit its own vertically integrated downstream business—while raising costs for competitors.[194] Supply squeezes are similar, but involve the refusal to sell or to prioritize sale of goods during a time of shortage.[195] In the AI context, vertical integration—or partnerships—across the technology stack raises the possibility of anticompetitive squeezes. Apple's partnerships with TSMC have raised questions about the semiconductor manufacturer preferencing Apple over other consumers of chips.[196] Cloud providers might charge advantageous rates to their affiliated foundation models and applications. Foundation model developers could charge different rates to their affiliated applications compared to their competitors. The market structure of the sector makes these real possibilities. Here too, however, doctrine has developed in a way that might make such claims difficult to win. In an important broadband internet case, the Supreme Court considered AT&T's integrated digital subscriber line (DSL) and internet service provider (ISP) businesses.[197] AT&T's rivals in the ISP business argued that it sold them wholesale DSL service at high prices, while selling its own retail service at a low price.[198] This made it impossible for these would-be rivals to effectively compete. The Court rejected the rival ISPs' claim, declaring that AT&T had "no duty to deal in the wholesale market," and thus no obligation to treat competitors on a level playing field with its own business line.[199]

Related to the refusal to deal is the essential facilities doctrine. Under the doctrine, a firm that controls an "essential facility" must give reasonable access to users, even if competitors. The doctrine requires that the essential facility is a monopoly, that it is infeasible for a competitor to replicate the facility, that the competitor has been denied access to the facility, and that the facility could offer access to the competitor.[200] Critical parts of the AI stack could be deemed essential facilities. Cloud infrastructure, data, and foundation models are all infeasible to duplicate for most businesses, due to their high costs to develop. These gives firms in these areas considerable power—and the potential to deny utility-like services to users. Indeed, the essential facilities doctrine could be seen as an antitrust remedy that seeks to implement NPU principles. But even as scholars have recently argued to extend the essential facilities doctrine to encompass technology and internet platforms,[201] it has remained disfavored by leading antitrust experts.[202] The

[194] HERBERT HOVENKAMP, PRINCIPLES OF ANTITRUST 293 (2017).

[195] *Id.*

[196] *See supra* TAN xx.

[197] Pacific Bell Tel. Co. v. LinkLine Communic., Inc., 555 U.S. 438 (2009).

[198] *Id.*

[199] *Id.* at 450-51.

[200] MCI v. AT&T, 708 F.2d 1081 (7th Cir. 1983).

[201] Nikolas Guggenberger, *Essential Platforms*, 24 STAN. TECH. L. REV. 237 (2021).

[202] Phillip Areeda, *Essential Facilities: An Epithet in Need of Limiting Principles*, 58 ANTITRUST L.J. 841 (1989); HOVENKAMP, PRINCIPLES, *supra*, at 297 (noting that it is "one of the most troublesome, incoherent and unmanageable" pathways to antitrust liability).

Supreme Court has also been skeptical of the doctrine. In *Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, LLP*,[203] the Court observed that "Compelling [infrastructural] firms to share the source of their advantage is in some tension with the underlying purpose of antitrust law, since it may lessen the incentive for the monopolist, the rival, or both to invest in those economically beneficial facilities."[204] While the Court did not completely reject the essential facilities doctrine, it also did not adopt it.[205]

Moreover, antitrust enforcement suffers from a number of other problems, as compared to other antimonopoly governance strategies. It operates *ex post*, with the Justice Department or FTC bringing cases to address restraints on trade or monopolization. The ex post nature of antitrust enforcement undermines effective governance in the AI sector because it allows consolidation to accrue and abusive practices to take place—potentially for years—before they may be redressed. The downsides to waiting are significant: Consolidation that takes place can reshape the market in ways that cannot effectively be undone later, or that are extremely resistant to change, due to network effects, and lock-in, among other market dynamics noted above. By contrast, ex ante governance rules are market shaping tools. They structure the market in favor of competition from the start, rather than trying to rework it once entrenched players have dominance or undertake bad behaviors.

It is also a case-specific process, in which mergers and other anticompetitive behaviors are addressed individually. This gives rise to a similar problem: Antitrust enforcers have to bring individual cases against every actor in the sector engaged in anticompetitive behavior. This can take considerable time, scaling up enforcement would be extremely resource-intensive, and the agencies themselves are resource constrained. The alternative to public enforcement—private enforcement—depends upon having one provider in this concentrated and interconnected network sue another. But these providers may have private incentives to avoid upsetting one another: one's application may depend on another's model; or one's cloud computing platform may depend on purchasing hardware from another.

And, in antitrust, a great deal of decisional power rests with courts rather than federal agencies. This is problematic for the many standard reasons that agency regulation is superior to court adjudication: Courts may be unpredictable and judges have little expertise in new technologies, especially as compared to legislators or agency experts. Judicial decisions, enshrined in precedent, are also less flexible to changes across time or context. Agencies, in contrast, are better able to take account of a broader set of facts and perspectives when crafting rules to drive firm behavior and can design rules for different situations.[206] And while

[203] 540 U.S. 398 (2004)

[204] *Id*. at 407–08.

[205] *Id*. at 410–11.

[206] For discussions of the downsides to a court-centric approach to antitrust, see Rebecca Haw, *Amicus Briefs and the Sherman Act: Why Antitrust Needs a New Deal*, 89 TEX. L. REV. 1247 (2011); Ganesh Sitaraman, *Taking Antitrust Away from the Courts*, GREAT DEMOCRACY INITIATIVE (2018). See also Anton Korinek, *Integrating Ethical Values and Economic Value*, in THE OXFORD HANDBOOK OF ETHICS OF AI 475, 491 (2020) (lauding a more participatory approach, explaining

courts are (by design) insulated from political accountability to the general public, agency governance is more democratic, as it incorporates public participation and is more responsive to political changes and popular opinion. In the case of AI, each these factors shows the benefits of ex ante governance.

For all these reasons, while antitrust law and enforcement remains important for shaping and policing markets, it will likely prove insufficient to address the urgency and scope of antimonopoly harms and practices related to AI. Layers in and across the AI tech stack are, as noted, structurally inclined towards consolidation and concentration, meaning that underenforcement of competition harms threatens to amplify the risks we have outlined above. Ex ante governance tools—described in Part IV—are likely to be essential to prevent these harms.

C. *The Benefits of Regulation for Innovation*

One common objection to the regulation of technological markets is that such regulations can harm innovation.[207] Our analysis of the AI technology stack not only undermines about this trope; it affirms a case for regulation as innovation-enhancing. The AI sector is currently subject to considerable market concentration at critical junctions in the technology stack and is vertically integrated across different layers as well. Such a structure is likely, over time, to result in less innovation than a more competitive market structure.

The reason, as we have seen, is that vertically-integrated firms that dominate utility-like services (such as cloud computing) can leverage that power in in downstream markets. This can happen through a variety of means: tying products, integrating products together, predatorily pricing competitors in downstream markets, charging unreasonable prices for utility-services to downstream competitors, copying the products of downstream competitors, and self-preferencing their own downstream products, among others. These practices can lead to less innovation overall because they restrict the diversity of the downstream product market first by pushing competitors out and second by then chilling investment and entry into the downstream market. Indeed, economists have modeled how technology platforms have created "kill zones" in which venture capitalists do not want to fund new startups because they know they will not be able to capitalize on their investment, because the tech platform will crush the new entrant.[208] For those who are

that "we need a large and concerted public effort . . . to ensure we develop AI in a direction that is both economically beneficial and ethically desirable.").

[207] Andrea O'Sullivan & Adam Thierer, *Counterpoint: Regulators Should Allow the Greatest Space for AI Innovation*, 61 COMMC'NS ACM 33, 33 (2018) ("[A]rtificial intelligence technologies should largely be governed by a policy regime of permissionless innovation so that humanity can best extract all of the opportunities and benefits they promise."); Andrew Stirling, *Precaution in the Governance of Technology*, *in* THE OXFORD HANDBOOK OF LAW, REGULATION, AND TECHNOLOGY 573, 577–78 (Roger Brownsword, Eloise Scotford & Karen Yeung eds., 2017) Thierer, *supra* note xx; Andrea O'Sullivan, *Don't Let Regulators Ruin AI*, MIT TECH. REV., Oct. 24, 2017, https://www.technologyreview.com/2017/10/24/3937/dont-let-regulators-ruin-ai/ (arguing that proposals for a new regulatory agency are based on the precautionary principle and are undesirable because they will limit innovation); *see also* JULIE E. COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM 90–92 (2019) (summarizing—and ultimately dismissing—such arguments).

[208] Kamepalli, Rajan & Zingales, *supra* note xx.

concerned about the beneficial uses of AI applications, this should be particularly troubling as it likely means fewer innovative applications will be developed that could have beneficial uses.

Regulation and other government policies can solve this problem and help spur innovation in downstream markets. The internet is one classic example: nondiscrimination and separations regulations between the providers of internet transmission services (i.e., bandwidth) and internet applications helped to foster exceptional growth in the latter market.[209] The airlines are another: There, regulation created a level playing field that ended destructive competition between firms and enabled investment in the high-capital-cost industry.[210] By preventing vertical integration and monopolization of an entire supply chain, NPU tools like structural separations and nondiscrimination rules (which we describe in more detail in Part IV) enable innovation at different points in that chain. Indeed, as we note in that Part, these tools were designed for traditional utilities—critical inputs into widespread applications. Applying these governance strategies to AI is also likely to create a more stable, predictable regime for competitors at all layers in the stack than a non-regulatory approach.

D. The Importance of Governing Market Structure

Finally, our analysis of the AI technology stack shows the importance of governing market structure, not just the particular harms from the conduct of an AI application, such as biased output or misinformation (what we call a "conduct harm"). For those who are in the technology industry and seek to start new companies, invest in them, or work for them, issues of market structure and dominance are critically important. Moreover, to the extent one thinks economic equality, resilience, and democracy are desirable, concentration in AI is again relevant. In short, we think the structural approach is an essential complement to conduct-based regulations.

Perhaps more surprisingly, a structural approach can also help address conduct harms—and in some cases might resolve them better than focus on application conduct. This is for two reasons. First, understanding the structure of AI's technology stack allows us to identify locations within the stack in which regulatory interventions might be most helpful for addressing downstream conduct issues. For example, if a concern stems from the quality of data that goes into train a model,[211] regulating data warehouses and data processing might be more valuable than focusing on AI models or applications. If the concern is the use of AI by bad actors, focusing on bottlenecks in AI's technology stack that all users rely on might be helpful. For example, placing liability for downstream uses on cloud providers or model providers could force them to develop systems to screen potential users;

[209] Tejas N. Narechania & Tim Wu, *Sender Side Transmission Rules*, 66 FED. COMMC'NS L.J. 467, 470–74 (2015).

[210] GANESH SITARAMAN, WHY FLYING IS MISERABLE AND HOW TO FIX IT (forthcoming 2023).

[211] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. 1, 6–11 (2018)

requiring licensing of all users at the cloud or model level could restrict users to those with training.[212]

Second, in some cases, confronting issues of market structure might help address conduct harms at the application level by increasing the diversity of options at the application level.[213] Imagine, for example, a cancer diagnostic software that has consolidated that market due to proprietary data lower down in the stack. The application provider could charge higher prices and slow its innovation to improve accuracy because it has no reasonable competitors.[214] Or consider a biased facial recognition application that gains dominance because it is part of a vertically-integrated AI company. Law enforcement using that application may not have many or any alternatives if the AI company has shuttered the competition—and controls the data and training capacity need to develop a workable model.

In other words, where any given application will tend to consolidate its control over an application market, due to vertical integration, or spectacularly high data and compute costs, the application provider is not only likely to tend towards higher user prices, it also gains a respite from the competitive pressures that typically force quality improvements. A more competitive market—even at these lower layers in the stack—might spur improvements in application accuracy and also help restrain prices. We should be clear about the scope of our claim: We do not mean to suggest competition is sure to address concerns about conduct harms like algorithmic bias and discrimination. Competition may help address these concerns through quality-based competition, and so we think that a more competitive environment is better for those who share these concerns. Competition regulation may be, in this regard, a critically important complement to regulations that seek to address concerns about bias, discrimination, and privacy, directly.[215]

IV. An Antimonopoly Approach for Artificial Intelligence

As we have seen, the AI technology stack is characterized by monopoly and oligopoly in specific layers. It is likely to remain so, due to features of the technology, and to become an oligopoly across the AI stack as well. These conclusions about the industrial organization of the AI sector suggest that ex ante governance tools are likely to be more effective than ex post tools in preventing anticompetitive behaviors and that they can also spur innovation and help address conduct harms

[212] Of course, there might be tradeoffs with this approach, such as raising barriers to entry due to the obligation to engage in these compliance processes.

[213] *See* Tejas N. Narechania, *Machine Learning as a Natural Monopoly*, 107 Iowa L. Rev. 1543, 1592–95 (2022) ("many of the now well-known problems attending to machine learning-based applications might be understood as problems of market power")

[214] *Cf.* Solow-Niederman, *supra* note xx, at 641 & n.34 (quoting Casey Ross & Ike Swetlitz, *IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show*, STAT (July 25, 2018), https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf).

[215] *See* Anton Korinek, *Integrating Ethical Values and Economic Value, in* The Oxford Handbook of Ethics of Ai 475, 487 (2020) (noting the necessity of "pass[ing] regulation to compel innovators to take into account their adverse effects of society"). We note that there is voluminous literature on such matters; a literature we have only begun to scratch the surface of in the sources cited in notes 1–19, among other sources cited throughout this Article and elsewhere.

from AI applications. In this final Part, we outline the antimonopoly tools—industrial policy, NPU rules, public options, and cooperative governance—that can help govern the structure of the artificial intelligence sector.

A. Industrial Policy and Industrial Organization

In the hardware layer, scarcity and supply chain vulnerability are paramount concerns. To address these problems, the United States has already taken steps to incentivize the development of chip manufacturing both within the United States and among a wider set of firms. The bipartisan Chips and Science Act of 2022[216] established a range of incentives to spur domestic production of cutting-edge chips. The Act committed $52.7 billion to the Departments of Commerce and Defense and the National Science Foundation to support U.S. development of semiconductor programs.[217] The Commerce Department's Chips for America program seeks to use federal funds to crowd-in private sector investment in order to develop at least two large scale clusters for fabrication of chips.[218] Whereas the Chips and Science Act spurs domestic development, other policies have been designed to ensuring U.S. leadership and prevent concentration of production capacity in China. The Biden Administration has thus placed export control restrictions on sharing advanced semiconductor technologies with certain Chinese entities.[219] There are also reports that the Biden Administration is preparing restrictions on outbound investment going to Chinese technology firms.[220]

One of the central questions for industrial policy in the AI sector is whether investment decisions will entrench dominant players or facilitate competition. Subsidies, loan guarantees, or tax advantages directed toward dominant players may simply keep them in positions of leadership. In areas that have a tendency toward consolidation—due to economies of scale, network effects, high capital costs, and other factors—such policies could further extend their lead. But industrial policies could also be targeted at new, smaller, and innovative actors, in which case they can facilitate competition, rather than entrench market power.[221]

It is too early to tell at this point whether U.S. industrial policies will entrench power or increase competition, but government officials coordinating industrial policy efforts—such as semiconductor programs under the Chips and

[216] CHIPS Act of 2022, Pub. L. No. 117-167, Div. A, § 102, 136 Stat. 1372 (2022)

[217] *Id.*

[218] CHIPS FOR AMERICA, VISION FOR SUCCESS: COMMERCIAL FABRICATION FACILITIES 1, NAT'L INST. STANDARDS & TECH., Feb. 28, 2023, https://www.nist.gov/system/files/documents/2023/02/28/Vision_for_Success-Commercial_Fabrication_Facilities.pdf

[219] Michael Schuman, *Why Biden's Block on Chips to China is a Big Deal*, THE ATLANTIC, Oct. 25, 2022, https://www.theatlantic.com/international/archive/2022/10/biden-export-control-microchips-china/671848/.

[220] Jack Stone Truitt, *Biden Executive Order on Investments in China Faces Hurdles*, NIKKEI-ASIA, June 10, 2023 (01:30 JST).

[221] See Philippe Aghion, Jing Cai, Mathias Dewatripont, Luosha Du, Ann Harrison & Patrick Legros, *Industrial Policy and Competition*, AM. ECON. J.: MACROECONOMICS 7(4): 1-32 (2015).

Science Act—could consider market diversification and competition as a critical element in evaluating candidates for federal grants.[222]

B. Tools from NPU Law

Regulatory tools from the law of networks, platforms, and utilities have long been applied to industries that feature network effects, and functional or actual monopoly or oligopoly characteristics.[223] NPU regulations, as a recent textbook describes, provide a legal framework that can help build NPUs at scale, ensure continuity of service, prevent monopoly and oligopoly abuses, avoid destructive competition, ensure widespread access, promote commercial development, and sustain democracy.[224] These regulations operate primarily *ex ante*, that is, by structuring the market (often to favor greater competition), identifying likely harms, and establishing rules to prevent those harms *before* they arise. In this subsection, we describe how selected NPU tools could be helpful to addressing the downsides of an AI oligopoly.

1. Structural Separations — Structural separations "limit the lines of business in which a firm can engage."[225] The central benefit of structural separations is that they prevent a business from self-preferencing or leveraging their power from one business-line into another. For example, under the Hepburn Act of 1906, railroads were prevented from carrying commodities from any company in which they also had a stake.[226] The idea behind the rule was that railroads should offer equal services to all shippers, rather than preferencing their own vertically-integrated shipping interests. In addition to preventing conflicts of interest and leveraging of profits, structural separations also limit the concentration of economic power and promote a diverse business ecosystem of users of the platform.[227] Perhaps most importantly, they can be more administrable than other policies, such as nondiscrimination rules (discussed in the next section). If a company is involved in the prohibited business line, it violates the rule. This is a far clearer rule than one that requires monitoring specific behaviors.

With respect to AI, there are number of places where structural separations could be useful.[228] Perhaps most notably, structurally separating the cloud layer from higher layers in the stack could address a wide range of market dominance problems identified above. It would treat cloud computing platforms as utility providers of a commodity product (namely, computational capacity) that is open for all kinds of uses—like electricity—and ensure that those providers cannot

[222] Note that the Chips office does appear to want *two* clusters in the United States, but does not commit to those being run by two independent firms. *See* CHIPS FOR AMERICA, VISION FOR SUCCESS, *supra*.

[223] RICKS ET AL., *supra* note xx, at 8-10.

[224] ID. at 11-21.

[225] ID. at 28

[226] Pub. L. 59-337 (June 29, 1906).

[227] For a discussion of this example and others, including a theory of structural separations, see Lina M. Khan, *The Separation of Platforms and Commerce*, 119 COLUM. L. REV. 973 (2019).

[228] *Cf.* William P. Rogerson & Howard Shelanski, *Antitrust Enforcement, Regulation, and Digital Platforms*, 168 U. PA. L. REV. 1911, 1934–36 (2020).

prioritize their own downstream business lines over their competitors'. Separation would likely also spur cloud providers to innovate on their cloud offerings, rather than on innovation that comes from vertical-integration.[229] This would, in turn, also facilitate innovation in the downstream markets where cloud users could develop a range of products and services, rather than being pushed into the cloud company's system.[230]

2. Nondiscrimination, Open Access, and Rate Regulation — One alternative to structural separation requirements are nondiscrimination and equal access rules, sometimes coupled with rate regulation.[231] Nondiscrimination rules allow a firm to operate two or more vertically-linked business lines, but require the firm to treat downstream businesses neutrally—including its own vertically-integrated business lines.[232] Nondiscrimination and equal access rules apply to both access and pricing. Most platforms have to be open to all comers who seek to use them, with limited exceptions.[233] All users must also be treated similarly in terms of price.[234] Historically, nondiscriminatory pricing rules required firms to file their prices, called "tariffs," and make them publicly available.[235] Transparency about prices and prohibitions on charging prices that diverged from the posted tariff ensured equal rates for customers. Equal pricing rules are an essential corollary to open access because firms could charge prohibitive prices as a workaround to evade nondiscriminatory access requirements.[236] In some cases, regulators have also directly set the rates firms can charge. Rate setting "is usually directed toward preventing NPU enterprises from lowering output and raising prices," while simultaneously ensuring a reasonable return on invested capital.[237]

Nondiscrimination and equal access rules complement structural separations in areas in which a business has market dominance or acts as a platform for downstream activity. The reason is that a structurally separated platform could still pick and choose its users or charge differential prices or prohibitively high prices—even if it is not self-preferencing its own vertically-integrated businesses. Nondiscrimination and equal access rules can be implemented on their own, but they may be a second-best strategy for addressing self-preferencing concerns because of administrability issues. In theory, neutrality between business lines should prevent self-preferencing. But in practice, it is more difficult for regulators to police and

[229] If regulators were to determine that integration of chips and cloud is desirable for effective service provision, then separating chips/cloud from higher levels in the stack would encourage innovation across both layers together—while preserving the innovative potential of competition further up the stack.

[230] *See* Tom Slee, *The Incompatible Incentives of Private-Sector AI, in* THE OXFORD HANDBOOK OF ETHICS OF AI 107, 122.

[231] RICKS ET AL., *supra* note xx, at 24-26.

[232] ID. at 24, 26, 29.

[233] Ganesh Sitaraman, *Deplatforming*, YALE L.J. (forthcoming).

[234] RICKS ET AL., *supra* note xx, at 24.

[235] ID. at 24.

[236] ID.

[237] ID. at 25.

enforce nondiscrimination rules, than structural separation requirements.[238] Regulators have to monitor or audit specific business practices and identify violations of pricing or treatment—or, at a minimum, respond to complaints from businesses who might fear reporting the platforms upon which they depend to regulators. Structural separations, by contrast, are a prophylactic rule: they prevent any commingling of business lines, and thus are easily administered.

In the AI context, nondiscrimination and equal access rules could be adopted at multiple places in the stack. At the hardware level, given the scarcity of chips, fabricators and designers could be required to serve customers equally—at least until chip fabrication becomes more widely available. At the cloud level, cloud providers should treat all downstream businesses in a nondiscriminatory fashion, be open to all comers, and offer transparent, uniform, publicly available prices. Open source and non-open source, but commercially available, data warehouses and lakes could also be subject to nondiscrimination and equal access rules. This would enable many model developers to use the data to develop and train new models. Foundation models and APIs could also be subject to such rules, so that app developers can tweak those models to develop new products and services.

3. Interoperability Rules — Interoperability rules lower barriers to entry and thus stimulate competition by "allowing new competitors to share in existing investments" and "imposing sharing requirements on market participants."[239] In the telecommunications context, for example, policymakers changed the dynamics of entry into local telephone markets not only through open access rules, but also through interconnection mandates: By requiring that each telephone provider interconnect with another, no one provider could wield its network effects as a sword, effectively consolidating control over the entire market. A customer could choose any provider, and still reap the benefits of a network that spanned the entire market. Rules that required a one provider to transfer a user's phone number to a competing provider (and thus required that the providers work together on an interoperable number portability system) also facilitated competition among providers by reducing switching costs for users. Those rules targeted a notable lock-in effect: It is quite cumbersome to let all your contacts know you have a new phone number.

Such requirements could be applied in AI contexts, too. Recall that among of the drivers of consolidation in the model layer are barriers to data acquisition and data network effects. One type of interoperability rule would be to mandate data sharing through federated learning. Federated learning is a technical "approach to machine learning where a shared global model is trained across many participating clients that keep their training data locally."[240] Rules that require a federated learning approach among competitors may be attractive to policymakers seeking to induce competition while ensuring that no one application, vertically

[238] *See, e.g.*, Rory Van Loo, *In Defense of Breakups: Administering a "Radical" Remedy*, 105 CORNELL L. REV. 1955 (2020).

[239] Narechania, *supra* note xx, at 1555.

[240] See TensorFlow Federated: Machine Learning on Decentralized Data, TENSORFLOW, https://www.tensorflow.org/federated (describing one approach to federated learning).

integrated with the underlying model, uniquely benefits from improvements made through continuous or reinforcement learning.[241] Instead, the model's improvements are derived from all the applications that use it—and are shared among all of them, too. Such rules might likewise require companies to share tools and techniques for filtering personal information and deduplication, not only to enable federated learning, but also to improve outcomes and protect user privacy.[242] Such forms of AI development may help to undermine the consolidation-driving network effects of the data sublayer.

Likewise, policymakers might consider rules that improve interoperability among cloud platforms, easing transitions from one provider's system to another. The lack of interoperability, and the problems of switching, are a real concern. Technologists, for example, have proposed entire systems—"Sky Computing"—aimed at addressing the switching and interoperability costs associated with using different cloud providers.[243] As different providers of cloud computing services specialize—moving away from offering a pure commodity "compute" resource to more bespoke computing resources and incorporating specialized applications (or utilizing specialized hardware)—some applications developers have found it difficult to take advantage of specializations across different providers. A developer might wish, for example, to train a model on one cloud provider—but use a different one for inferential applications. Or they may wish to switch an application developed on an OpenAI model to now query a Google foundation model (or a foundation model by some new competitor). In that a case, a common API, across providers, can lower switching costs, yielder greater competition.[244] In all, interoperability among distinct providers can facilitate competition, giving rise to better outcomes for participants in the downstream model and application layers and ultimately for consumers.

4. Entry Restrictions and Licensing Requirements — Congress has often established entry restrictions and licensing requirements for firms or individuals operating in many sectors of the economy. Such rules limit entry into a sector to firms that have registered with an appropriate regulator or otherwise have approval from the government (often in the form of a license or certificate).[245]

These provisions are usually justified on one (or more) of three different grounds. First, entry restrictions can ensure safety and reliability. By placing

[241] This approach is distinct from the one adopted in Europe via Gaia-X, which predominantly regards federated data storage, for the purposes of complying data localization requirements (e.g., rules that certain personal data be housed in certain locales). By contrast, federated learning can describe an interoperable approach to training, in which multiple applications or users train a single, shared foundation model through an interoperable standard.

[242] Of course, such rules would have to confront issues that might emerge from the use of any personal or sensitive data and the challenges of reliable anonymization.

[243] Ion Stoica and Scott Shenker. 2021. From Cloud Computing to Sky Computing. In Workshop on Hot Topics in Operating Systems (HotOS '21), May 31-June 2, 2021, Ann Arbor, MI, USA. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3458336.3465302

[244] *Cf.* Google v. Oracle, 593 U.S. __ (2021).

[245] *See* RICKS ET AL., *supra* note 34, at 29-30.

conditions on entry into a sector, regulators can ensure that firms will operate safely and effectively. Airline pilots (and airlines themselves), for example, must be licensed. Likewise, nuclear power plants are licensed, in part, to ensure safe operation. Second, in some markets (particularly those typically characterized as natural monopolies or oligopolies) competition can lead to waste and ultimately deter capital investment.[246] In the railroad industry, for example, firms competed vigorously to build railroad tracks at a high cost—but fierce competition over price sent them into bankruptcy or merger. The result was wasted expense, abandoned rail lines, and eventual consolidation. Entry restriction can prevent these downsides, creating a stable environment for capital investment. And, third, in sectors where universal service—i.e., ensuring that *everyone* can access the regulated service—is a critical policy goal, regulators will often limit entry to the market.[247] This is because open competition often undermines universal service policy goals. Some services, like energy provision, have costs that vary across geographies: urban centers are typically cheaper to serve (and hence are more profitable), while rural areas can be more expensive. Without entry restrictions and related regulations, providers will tend to compete to serve the cheaper and more profitable customers (with those customers enjoying the benefits of competition), while neglecting the more expensive customer base. But entry restrictions coupled with duty-to-serve rules can ensure that everyone has access to the regulated service, often at regulated rates (typically regulated by, in part, averaging the high-cost customers with the low-cost ones).

Such requirements might be applied to the AI technology stack at various layers. First, entry restrictions might be deployed to ensure that certain foundation models and their associated applications are effective, and do not pose substantial risks to health and safety, or of bias. Indeed, the FDA's process for approving medical systems that incorporate AI resembles this approach. Similarly, licensing rules could oblige cloud providers to "know their customers," as in banking law, and ensure that entities in the model layer have checks in place to ensure non-discriminatory access, fair pricing, and safety. Applications could also be required to register with the model or cloud they use, to make it easier to identify and address dangerous or problematic behavior on a *post hoc* basis. Likewise, entry restrictions might help to address concerns about costly and wasteful investment—and the tendencies towards consolidation—in the model layer, which are characterized by high fixed costs, scale economies, and network effects.

C. Public Options

Another policy tool for increasing competition and service reliability are public options. Public options are publicly-provided goods or services that coexist with private market options, offered at some (often regulatorily-)set price.[248] Public options can help ensure competition, as the public option disciplines private monopolists or oligopolists that might increase prices or reduce service quality.[249]

[246] ID.

[247] ID.

[248] SITARAMAN & ALSTOTT, *supra* note xx, at 27.

[249] ID. at 38-40

Competition from private parties, in return, ensures that the public provides high quality service as well.[250] A public option also adds to the diversity of the sources of production, even if slightly, thereby strengthening supply chain resilience and reliability.

In the AI context, a public option for cloud infrastructure could also serve as a helpful complement or alternative to structural separations or nondiscrimination and equal access rules.[251] Because of high capital costs, network effects, and concerns from vertical integration, a public option for cloud could provide the cloud services that developers and end-users need—but without relying on the oligopoly providers. The public option for cloud would increase competition, by offering an alternative to high-priced oligopoly providers. And it would ensure that cloud space is available at an affordable price to researchers and other users who might have different goals than private firms. Indeed, Japan is in the process of building a public option supercomputer, which will make cloud services available to companies focusing on AI.[252]

Notably, the National Science Foundation's proposal to offer a National AI Research Resource (NAIRR) has focused on public access to AI research. NAIRR seeks to "democratize access to AI resources" and therefore "must primarily be sustained through Federal investment."[253] However, the NSF's proposal is unclear on the degree to which NAIRR will be a public option, or whether government will contract with private companies for critical AI services.[254] It suggests NAIRR provide a mix of computational resources, including "commercial cloud" as an option.[255] It also suggests that NAIRR "include at least one large-scale machine-learning supercomputer" but then is unclear whether this would be a publicly-run resource.[256] Recently-introduced legislation to create a NAIRR suggests that it would offer "a mix of computational resources," including "on-premises, cloud-based, hybrid, and emergent resources," "public cloud providers providing access to popular computational and storage services," open source software, and

[250] E.S. Savas, *An Empirical Study of Competition in Municipal Service Delivery*, 37 PUB. ADMIN. REV. 717 (1977); E.S. Savas, *Intracity Competition between Public and Private Service Delivery*, 41 PUB. ADMIN. REV. 46 (1981).

[251] *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY 208–212 (2015).

[252] Nikkei Staff Writers, *Japan's METI to Build New Supercomputer to Help Develop AI at Home*, NIKKEI ASIA, July 24, 2023 (15:02 JST), https://asia.nikkei.com/Business/Technology/Japan-s-METI-to-build-new-supercomputer-to-help-develop-AI-at-home.

[253] NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE TASK FORCE, STRENGTHENING AND DEMOCRATIZING THE U.S. ARTIFICIAL INTELLIGENCE INNOVATION ECOSYSTEM: AN IMPLEMENTATION PLAN FOR A NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH RESOURCE 22 (Jan. 2023)

[254] *See* AI Now Institute & Data & Society Research Institute, *Democratize AI? How the Proposed National AI Research Resource Falls Short*, AINOW (Oct. 5, 2021), https://ainowinstitute.org/publication/democratize-ai-how-the-proposed-national-ai-research-resource-falls-short.

[255] *Id*. at 31.

[256] *Id*. ("This could be made available by leveraging an existing supercomputer or newly procured through a competitive bid process managed by the Operating Entity in consultation with the Steering Committee and relevant advisory boards.").

APIs.[257] This structure may require some amount of non-oligopoly cloud provision, as the on-premises, cloud-based system provision is separate from the one that describes public cloud providers. The NAIRR, if funded, should ensure there is a true public option, rather than a government contract for researchers to purchase compute and other resources from cloud providers while in the process, further entrenching them.

The NAIRR legislation also includes provisions for data access,[258] and the federal government already has several other initiatives under consideration that are aimed at releasing public datasets to support model development.[259] Data is a resource that depends on extraordinary scale. More public options for data "would provide a pathway for start-ups and public-sector organizations to develop abilities and products that would compete with those of the tech giants," but without relying on their data.[260]

D. Cooperative Governance

Cooperatives are firms that are owned by consumers, workers, or producers. Cooperatives generally operate according to seven principles: open and voluntary membership; democratic member control; members' economic participation; autonomy and independence; education, training, and information; cooperation among cooperatives; and concern for community."[261] These principles have been summed up as three guiding ideas: the owners of the company are users, users control the company, and the purpose of the company is to benefit the users.[262] Today, many familiar U.S. companies are cooperatives, such as outdoor retailer REI, SunMaid Raisins, Land O'Lakes, State Farm Insurance, and ACE Hardware.[263]

One of the primary features of cooperatives is that they can subvert monopoly power.[264] In the early 19th century, as Henry Hansmann and Mariana Pargendler have shown, corporations in NPU industries—turnpikes, canals, railroads, banks—were legislatively chartered monopolies.[265] This structure gave rise to standard monopoly concerns, such as monopoly pricing. Legislators addressed these concerns by adopting restrictive corporate voting rights that placed power in

[257] Sec. 5603(b). *See* Press Release, Rep. Anna G. Eshoo, AI Caucus Leaders Introduce Bipartisan Bill to Expand Access to AI Research, July 28, 2023, https://eshoo.house.gov/media/press-releases/ai-caucus-leaders-introduce-bipartisan-bill-expand-access-ai-research

[258] *Id*.

[259] *See, e.g.*, AI.gov, *AI Researchers Portal: Data Resources*, https://www.ai.gov/ai-researchers-portal/data-resources/ (last visited Aug. 21, 2023).

[260] Gansky, Martin & Sitaraman, *supra* note xx.

[261] UNDERSTANDING THE SEVEN COOPERATIVE PRINCIPLES, NAT'L RURAL ELEC. COOP. ASS'N, https://www.electric.coop/seven-cooperative-principles%E2%80%8B/

[262] *See* BRUCE J. REYNOLDS, COMPARING COOPERATIVE PRINCIPLES OF THE U.S. DEPT. OF AGRICULTURE AND THE INT'L COOPERATIVE ALLIANCE 2 (U.S. Dep't of Agric. ed. 2014).

[263] Peter Molk, *The Puzzling Lack of Cooperatives*, 88 TUL. L. REV. 899, 900 (2014)

[264] Sandeep Vaheeson & Nathan Schneider, *Cooperative Enterprise as an Antimonopoly Strategy*, 124 PENN. ST. L. REV. 1 (2019)

[265] Henry Hansmann & Mariana Pargendler, *The Evolution of Shareholder Voting Rights: Separation of Ownership and Consumption*, 123 YALE L.J. 948, 951, 954-55 (2014)

the hands of consumer-owners.[266] This corporate governance regime effectively turned NPU monopolies into "consumer cooperatives," in which the primary users of the firm's service were also the owners, with the effect of directly addressing common concerns about monopoly pricing and service.[267] In the late 19th century, as capital became more available, general incorporation laws became widespread, and corporations grew to national scale, the antimonopoly toolkit changed. Antitrust law, federal NPU regulation, and cooperatives emerged as successors to corporate chartering in order to address the problems of monopoly control.[268] Cooperative governance, Hansmann and Pargendler observe, acted as an alternative to "both the costs of monopoly and the costs of rate regulation."[269]

Cooperatives are an antimonopoly tool because they "accomplish vertical integration" in a way that limits exploitative conduct.[270] In sectors with durable market power, dominant firms can raise prices, reduce output, or reduce the quality of service, thereby transferring wealth from suppliers or customers and to shareholders in the form of higher dividends or stock buybacks.[271] Cooperative governance shifts the incentives of management from distant shareholders toward users of the firm, with any excess profit going back to those same users. In infrastructural industries, including those with network effects, cooperatives might be particularly helpful—not only because the cooperative governance regime avoids the extraction of monopoly rents but also because it distributes wealth more equitably. Rather than concentrate wealth among the shareholders of a platform-business, cooperatives distribute wealth across the user-owners.

In the AI context, cooperative governance could be a particularly useful tool to not only address concentration and abuses of power, but also to govern AI in a manner that distributes wealth more equitably and that is more consistent with the goals and values of its users.[272] At the cloud layer, the federal government could support the creation of a cooperative research-focused cloud, owned and operated by nonprofits, government, and universities to ensure sufficient compute and storage power for research into innovative, safe uses of AI—and without a shareholder profit motive. The federal government could also support the creation of a cooperative cloud for private companies, in which firms could train and operate models, and share in the ownership of the cloud, without fear that one of the big platforms will take their ideas or raises prices for the utility services they provide. One might even imagine a cooperative model where one earns stakes in a model or application by contributing data to its development. These options for cooperatives in the cloud layer would help introduce competition between the cooperative and private

[266] *Id.*

[267] *Id.*

[268] *Id.* at 945-55.

[269] HENRY HANSMANN, THE OWNERSHIP OF ENTERPRISE 169-70 (1996).

[270] Molk, *supra* note xx, at 912.

[271] Vaheeson & Schneider, *supra* note xx, at 9.

[272] *Cf.* Tom Slee, *The Incompatible Incentives of Private-Sector AI, in* THE OXFORD HANDBOOK OF ETHICS OF AI 107, 122 (noting the successful and collaborative nature of Wikipedia, concluding that "something is working on Wikipedia that is not working at . . . Facebook or Amazon").

platforms, while simultaneously offering greater access to AI resources and distributing wealth more equitably.

CONCLUSION

Artificial intelligence has sparked considerable conversation and concern. Understanding the AI technology stack shows that the aspects of the AI industry are already a monopoly or oligopoly and that a dominant oligopoly is likely to emerge across the AI stack as a whole. This market structure comes with a number of drawbacks, including abuses of power, national security and resilience challenges, widening economic inequality, and political influence that can undermine democracy.

There are, however, a number of antimonopoly tools that can help address these problems. Tools from the law of networks, platforms, and utilities; public options; and cooperative governance can all help facilitate competition and combat inequality. Industrial policy can be designed in a way that encourages a more diverse ecosystem, rather than entrenching incumbents.

Technology leaders have sometimes operated on the mantra of "move fast and break things."[273] Political leaders have allowed that approach to define technology in the early 21st century. The result has been a governance failure that has led to concentration and a range of economic, social, and political problems.[274] As policymakers debate governing AI early in its technological lifecycle, antimonopoly tools must be part of the conversation.

[273] JONATHAN TAPLIN, MOVE FAST AND BREAK THINGS: HOW FACEBOOK, GOOGLE, AND AMAZON CORNERED CULTURE AND UNDERMINED DEMOCRACY (2017).

[274] For discussions, see ROGER MCNAMEE, ZUCKED: WAKING UP TO THE FACEBOOK CATASTROPHE (2019); SHOSHANNA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM (2018); JOSH HAWLEY, THE TYRANNY OF BIG TECH (2021); ROB REICH, MEHRAN SAHAMI & JEREMY M. WEINSTEIN, SYSTEM ERROR: WHERE BIG TECH WENT WRONG AND HOW WE CAN REBOOT (2021).

Vanderbilt
Policy Accelerator
for Political Economy & Regulation

VANDERBILT UNIVERSITY

vu.edu/vpa