YALE LAW & POLICY REVIEW

An Antimonopoly Approach to Governing Artificial Intelligence

*Tejas N. Narechania & Ganesh Sitaraman**

Since OpenAI released ChatGPT, debates over the regulation of artificial intelligence (AI) 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 highly concentrated.

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, evinces market concentration at and across a number of layers. And we argue that an unregulated 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.

* 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 Rebecca Haw Allensworth, Jack Goldsmith, Nikolas Guggenberger, Sonia Katyal, Sarah West Myers, Puja Ohlhaver, Sabeel Rahman, Morgan Ricks, Douglas Schmidt, Matt Stoller, Jules White, and Tim Wu for helpful conversations and suggestions and to Mathew Cha, Ramsay Eyre, and Lynn He for research assistance. Our thanks, also, to Courtney Perales and the Editors of the *Yale Law & Policy Review* for their careful edits and thoughtful suggestions.

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 at this early stage in its technological lifecycle, antimonopoly tools must be part of the conversation.

INTRODUCTION

Since OpenAI released ChatGPT, debates over the regulation of 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 an Executive Order on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."[2] The European Parliament passed the A.I. Act to regulate risky uses of AI technology.[3] Sam Altman—OpenAI's Chief Executive—has endorsed greater regulation of AI systems,[4] 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.[5] Scholars and commentators have discussed a wide range of problems with AI and proposed regulatory strategies to address those problems.[6] Notable books and articles cover algorithmic bias,[7] misinformation and

1. Off. of Sci. & Tech. Pol'y, *Blueprint for an AI Bill of Rights*, WHITE HOUSE, https://www.whitehouse.gov/ostp/ai-bill-of-rights [https://perma.cc/82E7-4H7X].
2. Exec. Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023).
3. 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 [https://perma.cc/KZW6-NY8W].
4. Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing*, N.Y. TIMES (May 16, 2023), https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html [https://perma.cc/SK88-993Z].
5. James Vincent, *Elon Musk and Top AI Researchers Call for Pause on 'Giant AI Experiments'*, VERGE (Mar. 29, 2023), https://www.theverge.com/2023/3/29/23661374/elon-musk-ai-researchers-pause-research-open-letter [https://perma.cc/AVB4-UUYD].
6. For an overview applying a range of existing legal principles to AI, see JACOB TURNER, ROBOT RULES: REGULATING ARTIFICIAL INTELLIGENCE (2018).
7. *E.g.*, SARA WACHTER-BOETTCHER, TECHNICALLY WRONG: SEXIST APPS, BIASED ALGORITHMS, AND OTHER THREATS OF TOXIC TECH (2017); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018); SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. 1 (2018); Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).

disinformation,[8] algorithmic collusion,[9] labor displacement,[10] legal personhood for AI,[11] liability rules,[12] common-law regulation,[13] explainability and transparency,[14] the FTC's regulatory powers over AI systems,[15] the right to contest AI determinations,[16] AI and the

8. *E.g.*, CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016); ALLIE FUNK, ADRIAN SHAHBAZ, & KIAN VESTEINSSON, FREEDOM ON THE NET 2023: THE REPRESSIVE POWER OF ARTIFICIAL INTELLIGENCE (2023).

9. *E.g.*, Ariel Ezrachi & Maurice E. Stucke, *Artificial Intelligence & Collusion: When Computers Inhibit Competition*, 2017 U. ILL. L. REV. 1775 (2017).

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

11. *E.g.*, Lawrence Solum, *Legal Personhood for Artificial Intelligence*, 70 N.C. L. REV. 1231 (1992).

12. *E.g.*, David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117 (2014).

13. *E.g.*, 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).

14. *E.g.*, 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).

15. *E.g.*, Michael Spiro, *The FTC and AI Governance: A Regulatory Proposal*, 10 SEATTLE J. TECH., ENV'T & INNOVATION L. 26 (2020).

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

administrative state,[17] AI and constitutional rights,[18] and AI's role in the use of international force,[19] among other concerns.[20]

For all the interest in regulating AI, there has been little discussion of AI's market structure.[21] This is surprising because parts of the AI supply chain (i.e., the "layers" in the "AI technology stack," to use the parlance of

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

18. *E.g.*, Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1875 (2020).

19. *E.g.*, 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).

20. *See, e.g.*, Ryan Calo, *Artificial Intelligence Policy: A Primer and a Roadmap*, 51 U.C. DAVIS L. REV. 399 (2017) (including data and privacy issues); Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353 (2016) (discussing institutional competence for regulation).

21. To the extent there has been such discussion, 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. One of the rare works to examine competition aspects of AI is C. Scott Hemphill, *Disruptive Incumbents: Platform Competition in an Age of Machine Learning*, 119 COLUM. L. REV. 1973, 1975-81 (2019). One more recent work is Daniel A. Crane, *Antitrust after the Coming Wave*, 99 N.Y.U. L. REV. 1187 (2024). One notable work on the AI supply chain is Jennifer Cobbe, Michael Veale & Jatinder Singh, *Understanding Accountability in Algorithmic Supply Chains*, 2023 PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1186. Some public advocacy organizations have begun to focus attention on this issue. *See, e.g.*, *AI in the Public Interest: Confronting the Monopoly Threat*, OPEN MKTS. INST. (Nov. 15, 2023), https://www.openmarketsinstitute.org/publications/report-ai-in-the-public-interest-confronting-the-monopoly-threat [https://perma.cc/U5VN-9G68]; Amba Kak & Sarah Myers West, *AI Now 2023 Landscape: Confronting Tech Power*, AI NOW INST. (Apr. 11, 2023), https://ainowinstitute.org/2023-landscape [https://perma.cc/X3FH-8JYL]. Policymakers are also just beginning to realize the competition-based threats to the future of AI. *See* Exec. Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023); Bureau of Competition & Off. of Tech., *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 [https://perma.cc/PAZ9-KCRB].

the sector) are monopolistic or oligopolistic.[22] Indeed, one of us has described how machine learning—the algorithmic foundation for AI applications—has natural monopoly characteristics, even under narrow economic definitions.[23] 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.[24]

In this Article, we make the case for an antimonopoly approach to governing artificial intelligence. We show that AI's industrial organization, rooted in AI's technological structure, evinces market concentration within and across a number of layers. It is true that the market for AI applications appears to be booming: It seems like every day, a new company announces that it is launching a new AI-based application. But a closer look reveals that a small oligopoly of large and well-entrenched players controls the technologies on which these applications are based. We argue that an unregulated AI oligopoly has undesirable economic, national security, social, and political consequences. Our analysis of AI's industrial organization—i.e., the structure of the firms and markets that compose AI's supply chain—leads to some important conclusions: (1) important elements of the AI sector are stable enough to invite regulation, notwithstanding ongoing technical development; (2) ex ante regulatory tools are likely to prove more effective than modes of ex post enforcement, as under antitrust law; (3) regulation can help facilitate more downstream innovation, and the current market structure may, in fact, inhibit innovation; (4) some of the most prominent worries about AI—such as bias and privacy—might themselves be partly the result of market-structure concerns; and (5) open-source development may be only an imperfect substitute for other competition in light of AI's industrial organization. In

22. 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.

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

24. 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).

light of these conclusions, we show how antimonopoly[25] market-shaping tools—including 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.[26] 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 ("Google Cloud"), and Microsoft Azure ("Azure")) dominating the marketplace.

The model layer is more complicated than the first two because it includes three sublayers (and even more within those sublayers): data, models, and model access. One primary input for an AI model is data, and so the model 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.[27] Models are what many think of as "AI." These models are the output of an algorithmic approach to analyzing and "learning"[28]

25. By antimonopoly tools, we mean a set of policy actions that address the economic, political, and social drawbacks of monopolies and oligopolies. This toolkit includes antitrust law and policy but is not limited to it.

26. *But see infra* note 49 (noting one source that has described parts of the stack).

27. 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 (Jul. 12, 2022), https://arxiv.org/abs/2108.07258 [https://perma.cc/VZZ2-RZ7M].

28. Narechania, *supra* note 23, at 1550 n.25, 1551 n.35 (2022) (discussing the use of terms such as "training" "learning" and "understanding" in the context of machine learning and artificial intelligence).

from the inputs that begin in the data sublayer. This "training"[29] process is expensive, and so models can be costly to develop. The third sublayer consists of modes of accessing these models—model hubs and application programming interfaces ("APIs"). While model hubs and APIs both offer access to a foundation model, they operate quite differently from each other. Model hubs are platforms that host foundation models. Developers can often download a foundation model, with its statistical details (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, application 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,[30] 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, as well as harms to downstream innovation, among other concerns.[31] 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

29. *Id*.

30. One of us tried it. The joke was not funny.

31. *See infra* Section II.A.

infrastructure and for the economy more broadly.[32] 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 but can also concentrate wealth in a small number of individuals—executives and shareholders.[33] 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.[34]

Our analysis of the AI technology stack and the downsides of an AI oligopoly yields five important conclusions, particularly in view of some of the prevailing tropes regarding regulating artificial intelligence. First, some commentators have worried that AI is moving too quickly for regulation.[35] 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 are 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 (though not all) of the harms we describe are the traditional subjects of antitrust law, antitrust enforcement is unlikely to be sufficient.[36] As we show, the AI technology 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

32. *See infra* Section II.B.

33. *See infra* Section II.C.

34. *See infra* Section II.D.

35. *See infra* Section III.A. For one example, see Amy Gibbons, *AI is Moving Too Fast to Regulate, Security Minister Warns*, TELEGRAPH (June 9, 2023), https://www.telegraph.co.uk/news/2023/06/09/security-minister-artificial-intelligence-regulation [https://perma.cc/S265-DNLM].

36. *See infra* Section III.B.

underenforcement more likely in a sector that seems structurally inclined towards concentration. Such underenforcement presents the risk that anticompetitive behaviors will persist and that market power, distributive, security, and democratic concerns will become more acute. None of this is to say that antitrust enforcement should be abandoned. Rather, it is to observe that there are benefits to adopting antimonopoly tools that operate ex ante—including industrial policy, the tools of NPU law, public options, and cooperative governance. These market-shaping tools can help to prevent 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.[37] This is contrary to the popular cliché that regulation hinders innovation. Vertical integration across the AI technology stack is likely to restrict the number of providers of services at downstream layers in the stack, reducing innovation and choice. Many antimonopoly tools, such as interoperability rules, are innovation-enhancing: They can help to create a level playing field for downstream businesses that rely on some foundational service. Hence, these tools have the effect of reducing bottlenecks and concomitant anticompetitive conduct, thereby boosting new innovative activity.

Fourth, 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 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 whom 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.

37. *See infra* Section III.C.

Fifth, we are skeptical of claims that open-source AI models will introduce competition in ways that completely address the downsides of an unregulated AI oligopoly dominated by the biggest technology companies. Discussions of "open-source" AI and competition often fail to appreciate some of the limits that are inherent to, or put upon, open-source models. Indeed, AI models have been released across a gradient of openness, rather than on a binary. Moreover, our analysis of the AI technology stack shows that open-source models will not address the unregulated AI oligopoly at the hardware or cloud layers. Because model layer enterprises are dependent on these lower layers, concentration means that oligopolists in these layers can leverage their power downstream—through self-preferencing, tying, or integration vertically into their applications. Dominant platforms have used open-source systems in the past to entrench and maintain their power, and they may yet do so again. Hence, we are wary of claims that open-source development will solve the set of competition problems we describe.

In Part IV, we turn to more specific solutions. We outline how antimonopoly and competition tools—industrial policy, 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[38] 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 by addressing the power of dominant firms. We further note how government procurement rules could incorporate antimonopoly principles and tools. Second, NPU law has long governed sectors with tendencies toward monopoly and oligopoly. We show how various tools for regulating NPUs—structural separation requirements, interoperability mandates, nondiscrimination rules and open access requirements, as well as service and rate regulations—could be applied to the various layers in AI's

38. 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 INST. (July 30, 2019), https://rooseveltinstitute.org/publications/industrial-policy-and-planning [https://perma.cc/LWQ5-GM4M]; and Ganesh Sitaraman, *Industrial Revolutionaries*, AM. PROSPECT (Sept. 10, 2020), https://prospect.org/economy/industrial-revolutionaries-franklin-hamilton-madison-jackson [https://perma.cc/GAW4-2S89].

technology stack. Third, we argue that public options[39] could helpfully complement these other tools 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.[40]

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 debates over regulating AI. For example, President Biden's Executive Order on AI encourages federal agencies "to promote competition in AI and related

39. *See* GANESH SITARAMAN & ANNE ALSTOTT, THE PUBLIC OPTION (2019).

40. 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 it has not been framed around AI. *See* MAJORITY STAFF OF SUBCOMM. ON ANTITRUST, COM. & ADMIN. LAW OF THE H. COMM. ON THE JUDICIARY, 116TH CONG., INVESTIGATION OF COMPETITION IN DIGIT. MKTS. 109-120 (2020) [hereinafter 116TH CONG., INVESTIGATION OF COMPETITION], https://democrats-judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf [https://perma.cc/T3VQ-SEE8]. A few commentators, including one of us, have suggested a public option for AI, but only in popular writings and without much analysis. *See, e.g.*, Ben Gansky, Michael Martin & Ganesh Sitaraman, *Artificial Intelligence is Too Important to Leave to Google and Facebook Alone*, N.Y. TIMES (Nov. 10, 2019), https://www.nytimes.com/2019/11/10/opinion/artificial-intelligence-facebook-google.html [https://perma.cc/FEE8-6MG8]; Bruce Schneier & Nathan E. Sanders, *Build AI by the People, for the People*, FOREIGN POL'Y (June 12, 2023), https://foreignpolicy.com/2023/06/12/ai-regulation-technology-us-china-eu-governance [https://perma.cc/2GJK-25EU]. There is, of course, far more literature on antitrust enforcement, but again, this is usually framed around technology platforms generally, rather than AI specifically. *See, e.g.*, Erik Hovenkamp, *Platform Antitrust*, 44 J. CORP. L. 713 (2019). *But see supra* notes 21, 23 (noting several sources examining competition issues in AI contexts).

technologies," but provides little guidance beyond that exhortation.[41] Second, we make the case that antitrust enforcement is likely to be insufficient for governing AI's market-structure problems and advocate for a complementary 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[42] and perhaps indirectly to the law-and-political-economy ("LPE") movement.[43] NPU law has, until recently, lain fallow, with its legal tools often playing only a secondary role in some policy debates.[44] 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, resilience, distributive justice, and perhaps even democracy itself. Public policy must contend with questions beyond economic analysis, including the vast power and distributional concerns that might emerge from control of this technology.[45]

41. Exec. Order No. 14110, 88 Fed. Reg. 75191, 75208 (Nov. 1, 2023).

42. *See generally* MORGAN RICKS, GANESH SITARAMAN, SHELLEY WELTON & LEV MENAND, NETWORKS, PLATFORMS, AND UTILITIES: LAW AND POLICY (2022) (describing and analyzing NPU law). For an application of this body to technology 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).

43. *See generally* Jedediah Britton-Purdy, David Singh Grewal, Amy Kapczynski & K. Sabeel Rahman, *Building a Law-and-Political-Economy Framework: Beyond the Twentieth-Century Synthesis*, 129 YALE L.J. 1784 (2020) (looking beyond the "twentieth-century synthesis" in favor of a law-and-political-economy approach).

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

45. *See* DARON ACEMOGLU & SIMON JOHNSON, POWER AND PROGRESS (2023) (arguing that technological advancement often consolidates wealth and power but can also confer substantial social benefits if accompanied by intentional policy choices).

A few clarifications are also in order. First and foremost, we do not aim to address every potential problem with AI[46] 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,[47] or through notice-and-comment rulemaking under current law.[48] And any of these tools could be adopted (and 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 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.[49] AI's technology stack, which is visually described in Figure 1 below, consists of four primary 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),

46. For a helpful overview of many of the downstream, application-based problems, see Laura Weidinger et al., *Taxonomy of Risks Posed by Language Models*, *in* FACCT '22: PROCEEDINGS OF THE 2022 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 214 (2022), https://dl.acm.org/doi/pdf/10.1145/3531146.3533088?trk=public_post_comment-text [https://perma.cc/6P4G-P6U3].

47. *See generally* Ganesh Sitaraman & Morgan Ricks, *Tech Platforms and the Common Law of Carriers*, 73 DUKE L.J. 1037 (2024) (describing the common law of carriers and arguing that tech platforms should be subject to it).

48. *Cf.* Spiro, *supra* note 15, at 50-59 (arguing that the FTC should use its broad powers under Section 5 of the FTC Act to regulate AI).

49. 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, 655-57 (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.

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 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.[50]

50. 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?*, ANDREESEN HOROWITZ (Jan. 19, 2023), https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform [https://perma.cc/2KK4-CRPQ]; Brad Smith, *Governing AI: A Blueprint for Our Future*, TOOLS & 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 [https://perma.cc/2JF8-BRFW]; Sayash Kapoor & Arvind Narayanan, *Three Ideas for Regulating Generative AI*, AI SNAKE OIL (June 21, 2023), https://aisnakeoil.substack.com/p/three-ideas-for-regulating-generative [https://perma.cc/3X4N-RWBB]; Matt McIlwain, *Game On in the Generative AI Stack*, MADRONA (Mar. 20, 2023), https://www.madrona.com/game-on-in-the-generative-ai-stack/ [https://perma.cc/M8FG-T2W8]; Deedy Das (@deedydas), TWITTER (Mar. 16, 2023, 8:45 PM), https://twitter.com/debarghya_das/status/1636544140069711872 [https://perma.cc/ML2Z-MNBB]; *see also* Assaf Araki, *Demystifying the AI Infrastructure Stack*, INTEL CAP. (Apr. 2, 2020), https://www.intelcapital.com/demystifying-the-ai-infrastructure-stack [https://perma.cc/T3J7-63NM] (describing seven layers of the AI technical stack).

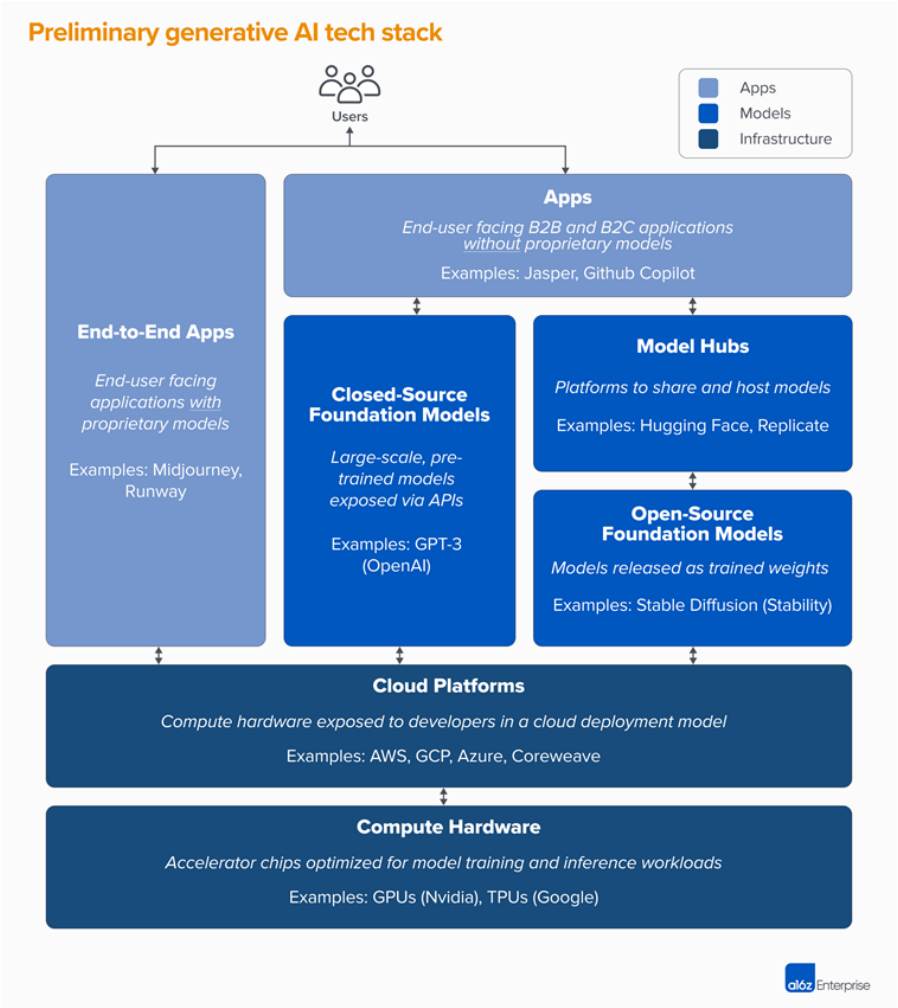

Figure 1. Source: Andreessen-Horowitz.[51]

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

51. Bornstein, Appenzeller & Casado, *supra* note 50.

number of calculations required by AI.[52] Chips come in three basic varieties. The primary of these are graphical processing units ("GPUs"), which were originally designed for processing images—a task that benefits from parallel (rather than sequential) computation.[53] 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.[54] 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 help deploy certain AI applications at scale.[55]

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. The trained algorithm then works through "prediction" or "inference," where the algorithm is deployed to engage a real-world scenario using its training. According to some estimates, "[h]undreds 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.[56] 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.[57] Microsoft might need more than 20,000 GPU servers, each with eight chips, to operate ChatGPT for all Bing users.[58] At a price of

52. Saif M. Khan & Alexander Mann, *AI Chips: What They Are and Why They Matter*, CTR. FOR SEC. & EMERGING TECH. 3 (Apr. 2020), https://cset.georgetown.edu/wp-content/uploads/AI-Chips%E2%80%94What-They-Are-and-Why-They-Matter-1.pdf [https://perma.cc/Z56L-DJXL].

53. *Id*. at 18.

54. *Id*.

55. *Id*. at 20-21.

56. 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 [https://perma.cc/Z5RN-KS8W].

57. Joe Lamming, *GPT-4: The Giant AI (LLaMA) is Already Out of the Bag*, VERDANTIX (Apr. 5, 2023), https://www.verdantix.com/insights/blogs/gpt-4-the-giant-ai-llama-is-already-out-of-the-bag [https://perma.cc/C4UP-ZWDU].

58. Leswing, *supra* note 56.

$10,000 a chip for 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.[59]

Each new generation of GPU accelerates AI development because microchips of prior generations seem to have increasingly "larger, slower, and more power-hungry transistors" and thus give rise to "huge energy consumption costs" that are "unaffordable" for all but the largest and most well-capitalized firms.[60] Nvidia's newer chip, the H100, was released in 2022 at a cost of $40,000.[61] Its performance is estimated to be three times better than its previous model.[62] Google has already built a supercomputer with 26,000 of the new GPUs.[63] Given the high energy costs, large technology companies often choose to physically locate their data operations close to sources of cheap electricity.[64]

The structure of the market for microprocessors is highly concentrated. As chip technologies have become 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.[65] Nvidia's chips are, in

59. *Id*.

60. Khan & Mann, *supra* note 52, at 6.

61. 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 [https://perma.cc/D3FN-RB4S].

62. *Id*.

63. 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 [https://perma.cc/J82V-PU4L].

64. *See* KATE CRAWFORD, ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE 215-16 (2021).

65. Leswing, *supra* note 56 (noting that Nvidia has 95% market share for machine learning); Zoe Corbyn & Ben Morris, *Nvidia: The Chip Maker that Became an AI Superpower*, BBC NEWS (May 30, 2023), https://www.bbc.com/news/business-65675027 [https://perma.cc/6MZM-8KUA] (same); Asa Fitch & Jiyoung Sohn, *The Next Challengers Joining Nvidia in the AI Chip Revolution*, WALL ST. J. (July 11, 2023), https://www.wsj.com/articles/the-next-challengers-joining-nvidia-in-the-ai-chip-revolution-e0055485 [https://perma.cc/96UT-UPJ2] (noting that Nvidia controls more than 80%

turn, manufactured (or "fabricated") by Taiwan Semiconductor Manufacturing Corporation ("TSMC"),[66] which is far and away the dominant semiconductor manufacturer.[67] Apart from TSMC, only Samsung fabricates similarly small, high-powered chips.[68] Making the smallest chips requires photolithography equipment, something only one company in the world, the Dutch firm ASML,[69] provides—and sells for up to $200 million per machine.[70] Figure 2, below, documents the growing concentration in this market over time.

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

Figure 2. Source: Center for Security and Emerging Technology.[71]

of the market); Wallace Witkowski, *Nvidia 'Should Have at Least 90%' of AI Chip Market with AMD on its Heels,* MARKETWATCH (July 11, 2023), https://www.marketwatch.com/story/nvidia-should-have-at-least-90-of-ai-chip-market-with-amd-on-its-heels-13d00bff [https://perma.cc/57KD-EFZE] (projecting that Nvidia will control 90% of the chip market).

66. Arjun Kharpal, *Two of the World's Most Critical Chip Firms Rally After Nvidia's 26% Share Price Surge*, CNBC (May 25, 2023), https://www.cnbc.com/2023/05/25/tsmc-asml-two-critical-chip-firms-rally-after-nvidias-earnings.html [https://perma.cc/2V5C-LYE9].

67. 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 (Mar. 13, 2023), https://www.eenewseurope.com/en/tsmc-globalfoundries-gained-as-foundry-market-cooled [https://perma.cc/G8DR-KL5M].

68. Khan & Mann, *supra* note 52, at 11.

69. *Id*. at 12.

70. 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), https://www.cnbc.com/2022/03/23/inside-asml-the-company-advanced-chipmakers-use-for-euv-lithography.html [https://perma.cc/B46A-EYAU].

71. Khan & Mann, *supra* note 52, at 12.

In recent years, large technology platform companies have entered the chip design business. Meta, for example, has designed a chip specifically for certain training and inference functions. As a Meta executive explained, "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."[72] Google, Amazon, and Microsoft have all likewise developed their own chips designed for specific AI-related functions.[73] Some of these chips, like Google's "Tensor Processing Unit," or TPU, are not general-purpose GPUs, but ASICs.[74] 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.[75]

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) requires significant processing power—sometimes called computational capacity or "compute"—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.

In general, cloud computing refers to the "ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources."[76] The "cloud" is simply hardware that exists somewhere else. It

72. Wiggers, *supra* note 63.

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

74. Nicole Kobie, *Nvidia and the Battle for the Future of AI Chips*, WIRED (June 17, 2021), https://www.wired.co.uk/article/nvidia-ai-chips [https://perma.cc/C8SR-PB85].

75. Khan & Mann, *supra* note 52, at 20-21.

76. PETER MELL & TIMOTHY GRANCE, NAT'L INST. STANDARDS & TECH., SPEC. PUBLICATION 800-145, THE NIST DEFINITION OF CLOUD COMPUTING: RECOMMENDATIONS OF THE

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 ("IaaS").[77] SaaS is the most familiar to the average consumer: It is the ability to run an application on one's own device as the application connects to the provider's remote servers or networks.[78] Google Docs is one example. PaaS 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."[79] IaaS provides users with "processing, storage, networks, and other fundamental computing resources."[80] While all three categories of cloud computing are relevant to AI, we focus here primarily on IaaS because both AI models and applications rely on infrastructural capabilities.

Cloud infrastructure features several dynamics that tend toward concentration and make sustaining competition difficult.[81] 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."[82] Second, there are significant costs to

---

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY 2 (Sept. 2011), https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf [https://perma.cc/SB23-BXA3].

77. *Id*. at 2-3.

78. *Id*. at 2.

79. *Id*. at 2-3.

80. *Id*. at 3.

81. *See Cloud Services Market Study: Final Report*, OFCOM 10 (Oct. 5, 2023), https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-3-4-weeks/244808-cloud-services-market-study/associated-documents/cloud-services-market-study-final-report.pdf [https://perma.cc/PD39-R2DU] (finding "clear indications that the cloud infrastructure market is not working well . . . [including] features of the market that we think have an adverse effect on competition and . . . [i]f left unchecked . . . could contribute to a further deterioration in competition in what is a critical market for digital services").

82. MAJORITY STAFF OF SUBCOMM. ON ANTITRUST, COM. & ADMIN. LAW OF THE H. COMM. ON THE JUDICIARY, 117TH CONG., INVESTIGATION OF COMPETITION IN DIGIT. MKTS. 96

shift from one provider to another. Some of these 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.[83] Some businesses have declared that even a twenty percent price discount is insufficient to overcome these concerns.[84] Such impediments to switching are exacerbated by additional costs imposed by cloud-computing providers, who sometimes charge "egress fees" on customers who take their data out of the cloud provider's system.[85] 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.[86] 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.[87]

Given these dynamics, the market for cloud-computing services has consolidated among three primary businesses (setting aside the long tail of smaller providers): AWS, Azure, and Google Cloud Platform ("GCP" or "Google Cloud").[88] The specific market shares of the firms vary by year and analyst but consistently suggest a strong degree of concentration, as Figure 3 below suggests. AWS is far and away the dominant provider, with more than 30% market share—and approaching 40% in some assessments. Azure comes in second near 20%, and Google Cloud and others run further

(2022) [117TH CONG., INVESTIGATION OF COMPETITION], https://www.govinfo.gov/content/pkg/CPRT-117HPRT47832/pdf/CPRT-117HPRT47832.pdf [https://perma.cc/W982-JMDS].

83. *Id*. at 269.

84. 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).

85. On egress fees imposed by AWS, see 117TH CONG., INVESTIGATION OF COMPETITION., *supra* note 82, at 269-70. For a discussion of the high costs of compute power, see Guido Appenzeller, Matt Bornstein & Martin Casado, *Navigating the High Cost of AI Compute*, ANDREESSEN HOROWITZ (Apr. 27, 2023), https://a16z.com/2023/04/27/navigating-the-high-cost-of-ai-compute [https://perma.cc/VJ4D-HXMR].

86. 117TH CONG., INVESTIGATION OF COMPETITION., *supra* note 82, at 31.

87. *See, e.g.*, *id*. at 269.

88. *Id*. at 93.

behind.[89] 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."[90] As a result, commentators have, as early as 1961, analogized cloud computing to other basic utilities.[91]

89. *See, e.g.*, Press Release, Gartner, Gartner Says Worldwide IaaS Public Cloud Services Market Grew 41.4% in 2021 (June 2, 2022), https://www.gartner.com/en/newsroom/press-releases/2022-06-02-gartner-says-worldwide-iaas-public-cloud-services-market-grew-41-percent-in-2021 [https://perma/cc/D8AU-4VBG]; Lionel Sujay Vailshery, *United States Cloud Infrastructure Services Vendor Market Share Q1 2021*, STATISTA (Dec. 6, 2022), https://www.statista.com/statistics/1237428/cloud-infrastructure-services-market-share-quarterly-us-vendor [https://perma.cc/5B67-MFT8].

90. Joe McKendrick, *Cloud Computing Market May Become an Oligopoly of High-Volume Vendors*, FORBES (July 11, 2013), https://www.forbes.com/sites/joemckendrick/2013/07/11/cloud-computing-market-may-become-an-oligopoly-of-high-volume-vendors [https://perma.cc/U3CW-XLVP].

91. SIMSON L. GARFINKEL, ARCHITECTS OF THE INFORMATION SOCIETY, THIRTY-FIVE YEARS OF THE LABORATORY FOR COMPUTER SCIENCE 1 (1999) (quoting John McCarthy speaking at the MIT Centennial in 1961, who commented that "computing may someday be organized as a public utility just as the telephone system is a public utility"); 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 [https://perma.cc/D29W-SDCG]; Rod Paddock, *The Cloud Networking Effect*, CODE MAG. (Dec. 16, 2021), https://www.codemag.com/article/1301011/The-Cloud-Networking-Effect [https://perma.cc/4FSW-7BFB] ("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?").

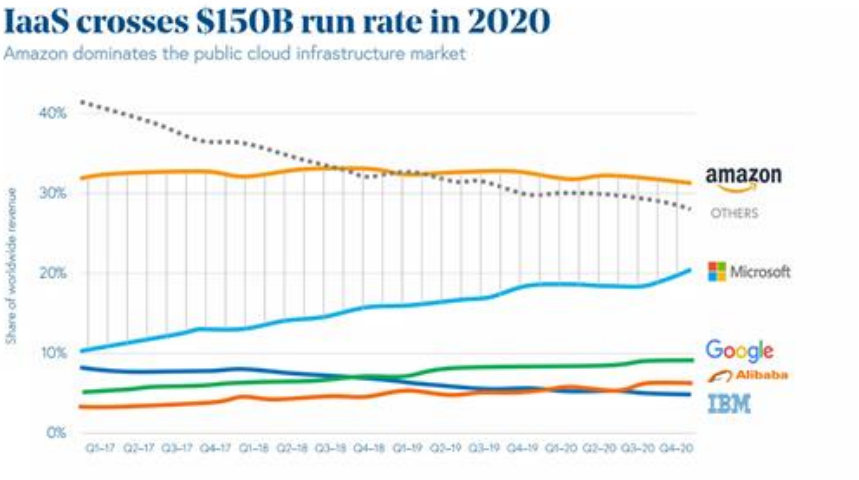

Figure 3. Source: Bessemer Venture Partners.[92]

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 ("LLMs"), such as ChatGPT, begin with extremely large corpora of text. GPT-3 is based on 300 billion "tokens" of text,[93] sampled from nearly 500 billion 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-

92. *State of the Cloud 2021*, BESSEMER VENTURE PARTNERS 21 (March 10, 2021), https://www.bvp.com/atlas/state-of-the-cloud-2021#Full-deck [https://perma.cc/6HAE-86N8] (demonstrating in slide 21 how companies have performed in the public cloud infrastructure market).

93. A token is a computational representation of text ranging from about four characters in the case of OpenAI's systems to as much as common two- or three-word phrases. *See Key Concepts*, OPEN AI, https://platform.openai.com/docs/concepts [https://perma.cc/HJ4K-BSGY].

language Wikipedia.[94] 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.[95]

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 a 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 comprises the model.[96] These basic, or "foundation," models, may, moreover, be tweaked or "fine-tuned" to particular purposes or applications.[97]

Downstream developers need to access the foundation models for fine-tuning and use in a particular application (e.g., GPT to generate text for a

94. Tom B. Brown et al., *Language Models Are Few-Shot Learners*, PROC. 34TH CONF. NEURAL INFO. PROCESSING SYS. (2020) (introducing and describing the GPT-3 language model). Some reports have put the training of GPT-3 at $10-12 million for each training run. Kyle Wiggers, *OpenAI's Massive GPT-3 Model Is Impressive, But Size Isn't Everything*, VENTURE BEAT (June 1, 2020), https://venturebeat.com/ai/ai-machine-learning-openai-gpt-3-size-isnt-everything [https://perma.cc/JXF6-46CK]; Alex Hern, *TechScape: Will Meta's Massive Leak Democratise AI—and at What Cost?*, GUARDIAN (March 7, 2023), https://www.theguardian.com/technology/2023/mar/07/techscape-meta-leak-llama-chatgpt-ai-crossroads [https://perma.cc/66R9-E2YA].

95. *GPT-4 Technical Report*, OPEN AI 2 (2023), https://arxiv.org/pdf/2303.08774.pdf [https://perma.cc/2R4H-CFCJ] ("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 compute, dataset construction, training method, or similar.").

96. 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 [https://perma.cc/B6C6-MXJK].

97. *See* Rishi Bommasani et al., On the Opportunities and Risks of Foundation Models 3 (Aug. 16, 2021) (unpublished manuscript), https://arxiv.org/abs/2108.07258 [https://perma.cc/HSY8-AHCK].

customer service chatbot). Some models, like OpenAI's, are closely held by their developer and accessible only via an API. An API is a "tool that 'allow[s] programmers to use . . . prewritten code to build certain functions into their own programs,"[98] 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.[99] The developer of a chatbot program might access GPT-3 through an API, using various commands to "fine-tune" the model for specific purposes,[100] and then send prompts to GPT and retrieve responses.[101] 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 consists of several layers, and this multilayer structure has important implications for market structure and competition—some of which are entangled with the concerns raised above regarding hardware and compute.

First is the data layer. As noted above, developing a model often depends on access to vast troves of data. In some instances, that data may be comparatively cheap to obtain. The repositories of internet text and Reddit posts used by ChatGPT, for example, are freely available for download.[102] Other providers have similarly scraped millions of publicly available images from online sources to train facial-recognition models.[103]

98. Google LLC v. Oracle America, Inc., 141 S. Ct. 1183, 1191 (2021) (quoting Oracle America, Inc. v. Google, Inc., 750 F.3d 1339, 1349 (2014)).

99. *Id*. at 1190.

100. *Fine Tuning*, OpenAI, https://platform.openai.com/docs/guides/fine-tuning [https://perma.cc/WKH3-5VAK].

101. *GPT Models*, OpenAI, https://platform.openai.com/docs/guides/gpt [https://perma.cc/BFP4-GCC9].

102. *About*, Common Crawl, https://commoncrawl.org/about [https://perma.cc/A629-ZZEL]; *Welcome!*, OpenWebText2, https://openwebtext2.readthedocs.io/en/latest [https://perma.cc/4KUB-7J87]; *see also* Brown et al., *supra* note 94, at 3-4 (describing the training data for GPT-3).

103. Olivia Solon, *Facial Recognition's 'Dirty Little Secret': Millions of Online Photos Scraped Without Consent*, NBC News (Mar. 17, 2019), https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921 [https://perma.cc/4J55-

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.[104] 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.[105] Google also paid $2.1 billion to purchase Fitbit, again with an eye toward collecting health metrics and data.[106] Even data that is easily collected or downloaded may present IP licensing costs, as a range of authors and creators have challenged model developers' claims that transforming these data inputs into AI models is "fair use."[107] Moreover, data may not be immediately usable for training an AI model. Data often require some combination of cleaning, validation, transformation, and labeling before being 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.[108] As training becomes more computationally complex, requirements in the processing and hardware layers increase dramatically because electrical power, processing, storage, and bandwidth requirements can grow polynomially or even exponentially.

HR3F]; 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-facial-recognition.html [https://perma.cc/JBY2-KRQR].

104. *See, e.g.*, Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 630 (2018) (arguing that well-designed access to copyrighted material could help AI developers reduce bias problems); C. Scott Hemphill, *Disruptive Incumbents: Platform Competition in an Age of Machine Learning*, 119 COLUM. L. REV. 1973, 1978-79 (2019).

105. Rob Copeland, *Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans*, WALL ST. J. (Nov. 11, 2019), https://www.wsj.com/articles/google-s-secret-project-nightingale-gathers-personal-health-data-on-millions-of-americans-11573496790 [https://perma.cc/X3FY-5YCS].

106. *Id*.

107. *E.g.*, Complaint at 4, N.Y. Times v. OpenAI, No. 1:23-cv-11195-SHS, 2024 WL 4301910 (S.D.N.Y. Dec. 27, 2023). These cases are still pending.

108. See Section I.A above for a discussion 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 [https://perma.cc/8MN5-KD4W].

These barriers to entry are significant: Developing a foundation model often requires substantial capital investment.[109] Some commentators have suggested that OpenAI was able to successfully develop GPT-3 only because it was "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."[110] As a result, the number of foundation models that can (or that perhaps should, from the standpoint of productive efficiency) exist for any given application class—language, image generation—may be quite limited.[111]

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).

The technical and market characteristics of the model layer highlight several ways in which it tends towards greater concentration. Start 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 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 rocket in popularity, previously free sources of training data are beginning to limit

109. Wiggers, *supra* note 94.

110. *Id*.

111. 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, at 4 (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), https://arstechnica.com/science/2020/12/why-ai-is-so-power-hungry [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").

access to this information or are now seeking to monetize it for AI training purposes.[112]

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 of us has written, 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 even more recent and relevant data that may further improve its application—and so on.[113]

Leading members of the industry have likewise observed that this process is a "virtuous circle for strengthening the best products and companies" and that AI thus appears as a "winner-take-all" system.[114] In short, scale can matter a lot to data,[115] and scale is becoming harder to achieve.[116]

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 a few foundation models are

112. *See* Mike Isaac, *Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems*, N.Y. TIMES (Apr. 18, 2023), https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html [https://perma.cc/VT7S-FFEZ].

113. Narechania, *supra* note 23, at 1584.

114. KAI-FU LEE, AI SUPERPOWERS: CHINA, SILICON VALLEY, AND THE NEW WORLD ORDER 19-20 (2018).

115. 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).

116. *See* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER 188 (2019) ("[M]achine learning is only as intelligent as the amount of data it has to train on, and Google has the most data.").

likely to emerge per application class. This gives model developers significant power that they may leverage into adjacent layers of the AI stack. Furthermore, some entities may leverage their scale in other layers to help them ensure and entrench dominance at the model layer.

For example, though LLaMA was likely very expensive to develop, Meta and Microsoft have decided to make that model public at no cost. Why? They are likely betting that providing 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.[117] This is in part because the costs of finetuning a model are minuscule compared to the costs of training it.[118] 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.[119]

Moreover, several 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

117. Steve Inskeep & Olivia Hampton, *Meta Leans on 'Wisdom of Crowds' in AI Model Release*, NPR (July 19, 2023), https://www.npr.org/2023/07/19/1188543421/metas-nick-clegg-on-the-companys-decision-to-offer-ai-tech-as-open-source-softwa [https://perma.cc/V2LY-2UUW] (quoting Nick Clegg, Meta's President for Global Affairs, 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").

118. *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 [https://perma.cc/58NC-UH5D].

119. 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), https://www.vice.com/en/article/5d3naz/openai-is-now-everything-it-promised-not-to-be-corporate-closed-source-and-for-profit [https://perma.cc/M957-YZ2E].

applications (such as Bard and Gemini, a competitor to ChatGPT).[120] 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 similar role: From chips and Azure cloud services to OpenAI's closed-source system of models, APIs, and applications, Microsoft has significant equities across the stack. 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.

As we elaborate below, 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, certain models, and APIs. But, developers relying on these resources may need to depend on dominant cloud operators to achieve scale. Indeed, some open-source resources expressly include limits ensuring they will not be used to challenge their proprietors.[121] 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."[122] 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

120. Wayne Ma, Anissa Gardizy & Jon Victor, *To Reduce AI Costs, Google Wants to Ditch Broadcom as its TPU Server Chip Supplier*, INFO. (Sept. 21, 2023), https://www.theinformation.com/articles/to-reduce-ai-costs-google-wants-to-ditch-broadcom-as-its-tpu-server-chip-supplier [https://perma.cc/J7JD-EMVX]; Janakiram MSV, *Google's Generative AI Stack: An In-Depth Analysis*, NEW STACK, (May 31, 2023), https://opendatascience.com/the-rapid-evolution-of-the-canonical-stack-for-machine-learning [https://perma.cc/GXY6-Y62W]; *see supra* Section I.B.

121. *LLaMA 3.1 Community License Agreement*, META (July 23, 2024), https://llama.meta.com/llama3_1/license [https://perma.cc/326K-FCR5] (explaining that LLaMA licensees cannot use the model to develop applications that serve more than 700 million users without Meta's approval).

122. David Gray Widder, Sarah West & Meredith Whittaker, *Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI*, NATURE (forthcoming 2024-25) (manuscript at 1) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4543807 [https://perma.cc/S3Z3-K6ZY].

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. Given these costs, it is possible that high-quality data resources and associated models will concentrate within a small number of dominant players.

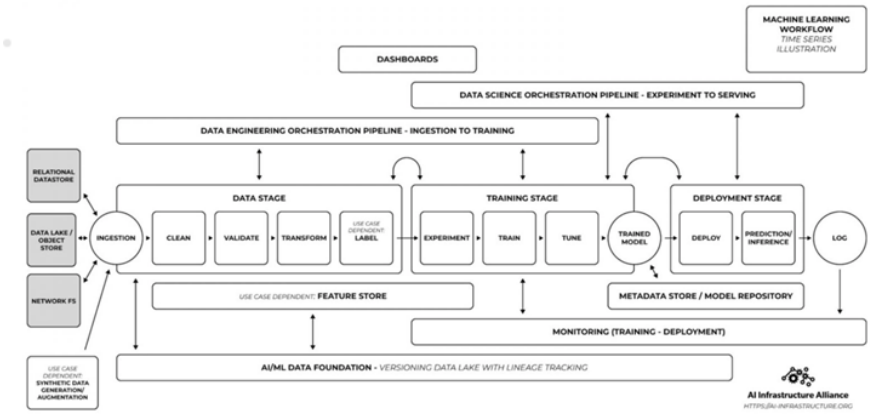

Figure 4. Source: AI Infrastructure Alliance.[123]

D. Applications

Finally, we come to the most visible layer in the AI technology stack—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 (say, "The application layer of the AI technology stack includes the user-facing tools and software applications that leverage AI models and algorithms to deliver

123. ODSC Community, *The Rapid Evolution of the Canonical Stack for Machine Learning*, OPEN DATA SCIENCE (July 13, 2021), https://opendatascience.com/the-rapid-evolution-of-the-canonical-stack-for-machine-learning [https://perma.cc/8RR4-M5GN]; *see also* Giancarlo Mori, *Demystifying the Modern AI Stack*, MEDIUM (Nov. 1, 2022), https://gcmori.medium.com/demystifying-the-modern-ai-stack-d91ce73ec4e [https://perma.cc/S7W5-38KK] (describing the "modern AI stack" as divisible into three internally complex parts: data management, model training/evaluation, and deployment).

specific functionalities, such as recommendations, language processing, or image recognition, directly to end-users.").[124]

The industrial organization of the application layer includes three categories. The first is 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.[125] In some cases, vertically integrated 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.[126]

The second category is 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.[127] Notably, the firms that operate the foundation models could themselves set up applications that compete with third-party developers and can set the terms through which data from the application is incorporated into future iterations of the underlying model.

The third category is 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

124. OpenAI, Response to "Describe the application layer of the AI technology stack in one sentence", CHATGPT (Nov. 5, 2024), https://chatgpt.com.

125. Frederic Lardinois, *Microsoft Launches the New Bing, with ChatGPT Built in*, TECHCRUNCH, (Feb 7, 2023), https://techcrunch.com/2023/02/07/microsoft-launches-the-new-bing-with-chatgpt-built-in [https://perma.cc/8MYA-B66Z]. 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 Corp. Blogs, *Microsoft and OpenAI Extend Partnership*, MICROSOFT: OFFICIAL BLOG (Jan. 23, 2023), https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership [https://perma.cc/M47S-7B8R].

126. Arti K. Rai, Isha Sharma & Christina Silcox, *Accountability, Secrecy, and Innovation in AI-Enabled Clinical Decision Software*, 7 J.L. & BIOSCIENCES 1, 3, 5 (2020).

127. *See, e.g.*, *Home*, GARDEN INTEL., https://www.gardenintel.com [https://perma.cc/6CQK-8CR5].

service chatbots.[128] 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 (calling on a model to resolve a particular query) is typically cheap—relatively speaking, especially when compared to the costs of model development and training.[129] But even if inference can seem relatively low-cost, inference at scale, which involves resolving thousands or millions of queries, still requires a substantial resource investment. This means that though developing an application might be comparatively cheap, scaling its use may prove to be very expensive.

* * *

In short, the AI stack consists of microprocessing hardware, models (data, models, and model hubs), and applications. This stack is already highly concentrated at its lower layers, and it is likely to remain that way given high capital costs, network effects, the difficulty of accessing sufficient data, and other barriers to entry. This concentration has significant drawbacks.

II. THE DRAWBACKS OF AN UNREGULATED 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 perhaps already are—dominated by a small number of firms.[130] Unregulated concentration in the AI sector—e.g., an unregulated AI oligopoly—has a variety of downsides. In this Part, we outline four sets of problems with an unregulated AI oligopoly: economic harms and abuses of power, national security and resilience issues, widening economic inequality, and effects on democracy.[131]

128. *See, e.g.*, *Home*, ADA, https://www.ada.cx [https://perma.cc/33KC-QM3R].

129. *See* Narechania, *supra* note 23, at 1580-81.

130. AJAY AGARWAL, JOSHUA GANS, & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE 216 (2018) ("For technology companies whose entire business might rest on AI, scale economies might result in a few dominant companies.").

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

A. Economic Harms and Abuses of Power

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, including concerns about price and quality, self-preferencing and discrimination, copying, and lock-in. Such anticompetitive conduct has a lengthy pedigree across a range of NPU sectors, including technology platforms.[132]

1. Price and Quality

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 one dominant firm in semiconductor manufacturing, it is quite possible that these firms could abuse their market power. They could, for example, demand monopoly prices for their goods, set discriminatory prices and terms for different customers, or refuse to deal entirely with some customers.[133] 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.[134] TSMC, on the other hand, has already prioritized its contracts and partnership with Apple over other chip consumers, giving lower priority to service, networking, and PC chips during periods of shortages.[135] Indeed, Apple has reportedly "locked up"

132. *See, e.g.*, RICKS ET AL., *supra* note 42, at 475-532 (discussing the origins and regulatory history of railroads); *id.* at 935-70 (discussing the origins and regulatory history of computer operating systems).

133. W. KIP VISCUSI, JOSEPH E. HARRINGTON, JR. & JOHN M. VERNON, ECONOMICS OF REGULATION AND ANTITRUST 82 (4th ed. 2005).

134. For a discussion, see generally RICKS ET AL., *supra* note 42 (discussing these abuses across a variety of industries, including railroads and operating systems).

135. Samuel Nyberg, *Apple Gets Special Treatment Amid Chip Shortage*, MACWORLD (June 22, 2021), https://www.macworld.com/article/677141/apple-gets-special-treatment-amid-chip-shortage.html [https://perma.cc/7NKJ-8UVB].

TSMC's entire capacity for fabricating 5-nanometer chips, which are currently the smallest and most advanced.[136]

Likewise, the concentration of cloud providers means high prices for users. Although the cost of compute has decreased over time[137]—cloud providers like AWS can charge substantial premiums (e.g., thirty percent margins) on the service.[138] 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 fifty public software companies could recover about $100 billion in market capitalization by switching to an in-house cloud platform.[139] 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, but once established, such companies should prefer proprietary service over these higher-cost external vendors. Nevertheless, they tend to stick with the higher-cost approach because of the costs and technical obstacles associated with switching.

Similar concerns also arise at the model layer, as concentrated control over foundation models gives rise to both price and quality concerns.[140] Foundation-model providers might, for example, deploy their market power to raise the costs to downstream developers for model access. Analogously, concentration among publicly available foundation models means reduced competition on the quality of the model—i.e., its value for use in downstream applications. Hence, the quality of each model matters

136. Jeremy Horwitz, *Apple Blamed IBM and Intel for Mac Chip Delays, but TSMC Won't Be Next*, VENTURE BEAT (Nov. 13, 2020), https://venturebeat.com/mobile/apple-blamed-ibm-and-intel-for-mac-chip-delays-but-tsmc-wont-be-next [https://perma.cc/5NNR-QNLX].

137. *See* Paddock, *supra* note 91.

138. Sarah Wang & Martin Casado, *The Cost of Cloud, a Trillion Dollar Paradox*, ANDREESSEN HOROWITZ (May 27, 2021), https://a16z.com/2021/05/27/cost-of-cloud-paradox-market-cap-cloud-lifecycle-scale-growth-repatriation-optimization [https://perma.cc/7X86-B4X2].

139. *Id*.

140. For a discussion of foundation models, including of competition concerns, see COMPETITION AND MKTS. AUTH., AI FOUNDATION MODELS: INITIAL REPORT, 2023, at 27-53 (UK), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1185508/Full_report_.pdf [https://perma.cc/P5ED-4E43].

more. If there are only one or two models in an application class (e.g., text, images) and those are flawed, then every application built on those models will suffer from those flaws.[141] The stakes for the quality of AI models are thus widespread.

2. Self-Preferencing and Discrimination

As we described above, providers have also vertically integrated across higher and lower layers of the technology stack, thereby raising a host of concerns and potentially complex questions.[142]

For example, the vertical integration across the AI stack may enable some providers to offer more—and more tailored—services. While more integrated offerings might seem beneficial, they also come with substantial downsides. One of the most significant such concerns regards these players' powers to favor vertically integrated AI-based applications, as self-preferencing has long precipitated competition concerns in network and platform industries.[143] Third-party firms rely on cloud services and model providers to develop their own proprietary applications—applications that compete with the cloud or model providers' own offerings. This can mean vulnerability to self-preferencing by that provider. Where a model provider offers an API to developers to create third-party applications (as OpenAI

141. 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 & Shmargaret Shmitchell, *On the Danger of Stochastic Parrots: Can Language Models Be Too Big?*, ASS'N FOR COMPUTING MACH., (Mar. 1, 2021) https://dl.acm.org/doi/pdf/10.1145/3442188.3445922 [https://perma.cc/S6BP-5KL2].

142. *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 become 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 in the market for services that exploit that infrastructure.").

143. For example, concerns have long been at the heart of related debates, such as network neutrality. *See, e.g.*, Protecting and Promoting the Open Internet, 80 Fed. Reg. 19738, 19749 (Apr. 13, 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.").

does at present), the model provider might also create its own competing applications, offering those affiliated applications favorable terms, such as charging higher rates to third-party developers than their own in-house business lines, early access to better versions of the model, or other advantages.[144] For example, if people ask Bing what they should do this weekend, it might suggest playing the video game Call of Duty—which Microsoft also owns.[145] At the extreme, model providers might exclude some third-party applications from using the model altogether, thereby preventing any competition with the affiliated application.

It is true that these concerns must be weighed against the possible benefits of vertical integration. That is, there may be some benefits to having a tightly integrated offering across layers. Consider, again, the dynamics of the GPU market. 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 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 less processing capacity (per inference)—and can benefit from ASICs if the tasks are repetitive and predictable.[146] 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

144. *See supra* notes 120, 129, and accompanying text.

145. We thank Nick Garcia of Public Knowledge for this example. We emphasize, moreover, that this integration is not limited to just these two layers. 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 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.

146. Andrej Karpathy, *Software 2.0*, MEDIUM (Nov. 11, 2017), https://karpathy.medium.com/software-2-0-a64152b37c35 [https://perma.cc/LH4P-3VEB] (noting that one benefit of neural networks, which are used for AI, is that they can be programmed into a chip).

may move down the stack, at least in part, in order to reduce their reliance on—vulnerability to—some of the dominant hardware providers.[147] And in response, it seems that chip designer Nvidia is also moving up the stack, offering its own cloud computing services.[148]

These developments invite an evaluation of three different possible market structures. 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

147. 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*, INFO. (Sept. 21, 2023), https://www.theinformation.com/articles/to-reduce-ai-costs-google-wants-to-ditch-broadcom-as-its-tpu-server-chip-supplier [https://perma.cc/J7JD-EMVX].

148. Press Release, 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 [https://perma.cc/VT84-UBBF]. Perhaps because they do not have cloud infrastructure built up, Nvidia is partnering or plans to partner with Oracle, Microsoft, and Google to provide this service, and is planning to buy 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*, INFO. (July 18, 2023), https://www.theinformation.com/articles/nvidia-accelerates-ai-startup-investments-nears-deal-with-cloud-provider-lambda-labs [https://perma.cc/FMV8-NRMS]; Anissa Gardizy, *In an Unsual Move, Nvidia Wants to Know Its Customers' Customers*, INFO. (July 31, 2023), https://www.theinformation.com/articles/in-an-unusual-move-nvidia-wants-to-know-its-customers-customers [https://perma.cc/TX8F-B4JJ]. 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*, INFO. (Sept. 11, 2023), https://www.theinformation.com/articles/nvidia-muscles-into-cloud-services-rankling-aws [https://perma.cc/32Q6-P6XJ].

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, as we have described above. The monopoly structure is concerning 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.[149] The vertically integrated structure also poses risks to competition because of the concerns noted above, regarding favoritism and foreclosure: Existing players might favor affiliates or have an incentive to lock out putative competitors.[150] And because hardware and cloud infrastructure would be integrated, new entrants would have to operate in both layers to compete effectively.

The 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.[151] Just as importantly, the fact that ASICs can be developed for specific inferential tasks suggests

149. 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 . . . ." Clay Pascal, *Nvidia H100 GPUs: Supply and Demand*, GPU UTILS (Nov. 2023), https://gpus.llm-utils.org/nvidia-h100-gpus-supply-and-demand [https://perma.cc/7MDB-RTWK].

150. Even skeptics of monopoly-leveraging theories (due, for example, to the one-monopoly-profit theory) might be persuaded of the possibility for leveraging in this context. JONATHAN E. NUECHTERLEIN & PHILIP J. WEISER, DIGITAL CROSSROADS: TELECOMMUNICATIONS LAW AND POLICY IN THE INTERNET AGE 14-17 (2d ed. 2013) (describing the theory and exceptions to it); Philip J. Weiser, *Toward a Next Generation Regulatory Strategy*, 35 LOY. U. 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.").

151. Nicole Kobie, *Nvidia and the Battle for the Future of AI Chips*, WIRED (June 17, 2021), https://www.wired.co.uk/article/nvidia-ai-chips [https://perma.cc/4K7U-RKGE].

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.[152] 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 to craft an appropriate response, drawing from the proposals we set out in Part IV.

3. Copying

The vertical integration we have described also raises concerns about copying, or what has sometimes been called "strip mining." Here, providers copy applications from downstream developers and incorporate them into their own offerings.[153]

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.[154] The prospect of expropriation of

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

153. Gerald Berk & AnnaLee Saxenian*, Rethinking Antitrust for the Cloud Era*, 51 POL. & SOC'Y 409, 416 (2023).

154. *Id.*; *see also* Andrew Leonard, *Amazon Has Gone from Neutral Platform to Cutthroat Competitor, Say Open Source Developers*, ONEZERO (Apr. 24, 2019), https://onezero.medium.com/open-source-betrayed-industry-leaders-accuse-amazon-of-playing-a-rigged-game-with-aws-67177bc748b7

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?[155] 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. Similar concerns attend to applications built upon foundation models. The model developers might appropriate the application and integrate its features into its own offerings. Venture capitalists call this the "kill zone," and leading economists have modeled its existence.[156]

4. Anticompetitive Acquisitions

Dominant firms have also employed a range of other tactics to prevent possible competitors from emerging, thus reducing innovation and competition overall. These include strategic acquisitions, which scholars explain have the effect of "coopting" putative disruptive entrants—to the benefit of existing dominant firms.[157]

[https://perma.cc/M27Y-9J3N]; Jordan Novet, *Amazon Steps Up Its Open-Source Game, and Elastic Stock Falls as a Result*, CNBC (Mar. 12, 2019), https://www.cnbc.com/2019/03/12/aws-open-source-move-sends-elastic-stock-down.html [https://perma.cc/RSZ4-5923]; 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 [https://perma.cc/4KDE-8VUD]. This provides two additional examples of Amazon introducing services that copy and compete with companies reliant on Amazon's cloud infrastructure.

155. *See* RICKS ET AL., *supra* note 42, at 15-16.

156. Sai Krishna Kamepalli, Raghuram Rajan & Luigi Zingales, *Kill Zone*, (Nat'l Bureau of Econ. Rsch., Working Paper No. 27146, 2022), https://www.nber.org/papers/w27146 [https://perma.cc/RE52-CVAA].

157. *See* Mark A. Lemley & Matthew T. Wansley, *Coopting Disruption*, 104 B.U. L. REV. (forthcoming) (manuscript at 16-36) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4713845 [https://perma.cc/8ZUU-MXRT] (discussing these tactics in addition to coopting venture capital and shaping regulation); Khan, *supra* note 24, at 755-90 (discussing leveraging, acquisitions, capital market dynamics, and other tactics).

5. Lock-In

Finally, dominant providers have taken steps to entrench their dominance, including by facilitating lock-in effects that raise the costs for consumers to switch providers.[158] In the context of cloud computing, for example, the lack of interoperability, the need for expertise in disparate systems, multi-year contracts, and egress fees all impede competition in the market.[159]

*

In all, the downstream effects of these abuses of power are potentially significant. For one, dominant providers can use their power to avoid both price and quality competition. Moreover, while there may be some reason to favor vertical integration in some limited instances, dominant and integrated providers can leverage their power in one layer across the other parts of the AI stack. Amazon, Microsoft, and Google are, as noted, developing microprocessing units specific to AI applications in order to fully integrate their hardware components across chips and cloud. They also offer applications that could integrate AI models. Putative applications developers may decline to develop new applications, if they believe incumbents will steal and copy their idea or will eventually lock them out of the market. This raises substantial concerns for would-be competitors. 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.[160]

158. *See* Fed. Trade Comm'n, No. 2023-0028-0001, Solicitation for Public Comments on the Business Practices of Cloud Computing Providers (March 22, 2023), https://www.regulations.gov/document/FTC-2023-0028-0001.

159. 116TH CONG., INVESTIGATION OF COMPETITION., *supra* note 40, at 98-99.

160. *Microsoft Threatens to Restrict Data from Rival AI Search Tools*, REUTERS (Mar. 26, 2023), https://www.reuters.com/technology/microsoft-threatens-restrict-data-rival-ai-search-tools-bloomberg-news-2023-03-25 [https://perma.cc/8SPR-UHF4].

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 foundry could be shut down due to a pandemic, weather event, war, or other emergency is—and, indeed, has been—significant.[161] Concentration leads to a fragile supply chain that is vulnerable to single points of failure across the stack.[162] More specifically, there are clear national security concerns with respect to the supply chain for chips.[163] Given that chips power not only AI but other critical technologies, the lack of availability could impede both military and non-military critical infrastructure.[164] 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.[165]

In addition to resiliency concerns, those focused on global competition and international leadership have observed that staying ahead on technology will be critical to global power in the twenty-first century.[166] In this context, the dominance of a single company in semiconductor manufacturing—and a company located in Taiwan—raises risks. A more geographically diverse supply chain with multiple firms (including U.S. production) would help ensure American leadership in cutting-edge technology.

---

161. *See, e.g.*, KAREN M. SUTTER, EMILY G. BLEVINS & ALICE B. GROSSMAN, CONG. RSCH. SERV., R47558, SEMICONDUCTORS AND THE CHIPS ACT: THE GLOBAL CONTEXT 2 (2023) (examining how U.S. policy efforts to promote U.S. semiconductor capabilities shapes and influences the global semiconductor industry).

162. On economic resilience as a critical part of foreign policy, see Ganesh Sitaraman, *A Grand Strategy of Resilience*, FOREIGN AFFAIRS (Aug. 11, 2020), https://www.foreignaffairs.com/articles/united-states/2020-08-11/grand-strategy-resilience [https://perma.cc/X6G3-DBQY].

163. For a history of the relationship between foreign policy and chips, see CHRIS MILLER, CHIP WAR: THE FIGHT FOR THE WORLD'S MOST CRITICAL TECHNOLOGY (2022).

164. *Id*. at 327-34.

165. *Id.* at 335-44.

166. Eric Schmidt, *Innovation Power: Why Technology Will Define the Future of Geopolitics*, FOREIGN AFFAIRS (Feb. 28, 2023), https://www.foreignaffairs.com/united-states/eric-schmidt-innovation-power-technology-geopolitics [https://perma.cc/GR59-FB96].

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 among them, gives rise to substantial software supply chain concerns.[167] 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.[168] Such disruptions would not only harm the affected companies but could have devastating effects on the economy as a whole.[169] Indeed, the interdependencies among cloud providers and downstream users are inscrutably complex, meaning that the chain reaction of outages would be difficult to predict without greater transparency requirements.[170] The lack of interoperability means that these systems could not easily be restarted on another provider's service. Hence, for the U.S. government and military, the 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 improper remedies. Perhaps, during normal times, regulatory processes and

167. *See* Press Release, U.S. Dep't of the Treasury, New Treasury Report Assesses Opportunities, Challenges Facing Financial Sector Cloud-Based Technology Adoption (Feb. 8, 2023), https://home.treasury.gov/news/press-releases/jy1252 [https://perma.cc/WYN7-K8SN] ("The current market is concentrated around a small number of CSPs [Cloud Service Provider], which means that if an incident occurs at one CSP, it could affect many financial sector clients concurrently.").

168. *See, e.g.*, Nick Merrill & Tejas N. Narechania, *Inside the Internet*, 73 DUKE L.J. ONLINE 35, 36-37 (2023).

169. Press Release, U.S. Dep't of the Treasury, New Treasury Report Assesses Opportunities, Challenges Facing Financial Sector Cloud-Based Technology Adoption (Feb. 8, 2023), https://home.treasury.gov/news/press-releases/jy1252 [https://perma.cc/WYN7-K8SN] ("Many financial institutions have expressed concern that a cyber vulnerability or incident at one [cloud service provider] may potentially have a cascading impact across the broader financial sector.").

170. *Cf.* Merrill & Narechania, *supra* note 168, at 62-66. (highlighting a similar problem among related but distinct providers of cloud-computing services).

protections, or competition, would suffice to catch these errors. But if only one firm has the capacity to deploy such technologies in an emergency (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 the 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.

Moreover, the standard economic harms from concentration, described above, can themselves give rise to national security implications. A lack of competition could mean lower levels of innovation over time, which could impact defense capabilities. An unregulated AI oligopoly will also focus primarily on economically profitable activities—activities that might not align with public goods or national security needs. And procurement contracts with AI firms for defense and national security purposes are likely to reflect the concentration in the industry, yielding high prices and quality problems.[171]

C. Economic Inequality

Concentration at layers within and across the AI technology stack can also deepen economic inequality in at least two 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, "Corporate profits will explode, showering wealth on the elite executives and engineers lucky enough to get in on the action."[172] For the United States, which is already on the high end of historic economic inequality in the population,[173] continuing the concentration of income and wealth both

171. For a discussion of these harms, see generally Ganesh Sitaraman & Alex Pascal, *The National Security Case for Public AI*, VAND. POL'Y ACCELERATOR FOR POL. ECON. & REGUL. (2024), https://cdn.vanderbilt.edu/vu-URL/wp-content/uploads/sites/412/2024/09/27201409/VPA-Paper-National-Security-Case-for-AI.pdf [https://perma.cc/JJ5D-UGRZ].

172. LEE, *supra* note 114, at 171; *see also* ZUBOFF, *supra* note 116, at 500 (noting "that GM employed more people during the height of the Great Depression than either Google or Facebook employs at their heights of market capitalization").

173. *See, e.g.*, THOMAS PIKETTY, CAPITAL IN THE TWENTY-FIRST CENTURY 291-303 (Arthur Goldhammer trans., 2014); Emmanuel Saez & Gabriel Zucman, *The Rise of*

arrests economic mobility[174] and is undesirable for those who seek a more egalitarian society.

Second, 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.[175] 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: Predominantly English-based systems accelerate the threats, for example, to endangered languages.[176] 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."[177]

Income and Wealth Inequality in America: Evidence from Distributional Macroeconomic Accounts, 34 J. ECON. PERSPS. 3, 5-17 (2020), https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.34.4.3 [https://perma.cc/67DN-6GCV].

174. For an overview of this argument and data, see Jared Bernstein & Ben Spielberg, *Inequality Matters*, ATLANTIC (June 5, 2015), https://www.theatlantic.com/business/archive/2015/06/what-matters-inequality-or-opportuniy/393272 [https://perma.cc/5JTL-8R7A].

175. Steven Weber & Gabriel Nicholas, *Data, Rivalry and Government Power: Machine Learning Is Changing Everything*, 14 GLOB. ASIA 22, 23-26 (2019), https://globalasia.org/data/file/articles/f95045850aa30d155ee4d75911d2c7a1.pdf [https://perma.cc/W5P3-8WF5].

176. Viorica Marian, Opinion, *AI Could Cause a Mass-Extinction of Languages—and Ways of Thinking*, WASH. POST (Apr. 19, 2023), https://www.washingtonpost.com/opinions/2023/04/19/ai-chatgpt-language-extinction [https://perma.cc/G9ER-62X7]. 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).

177. LEE, *supra* note 114, at 172.

D. Democracy

The concentration of economic power has long been understood as a danger to a republican form of government.[178] In the AI context, concentration in and across the technology stack raises concerns for the health of our democracy.[179] For starters, democracy depends on vibrant political debate and discussion.[180] 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 of 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 is an issue because private firms are guided by private interests, rather than the public good, and thus may have an interest in facilitating information that is financially beneficial even if otherwise problematic. The latter is an issue 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, therefore, to democratic government.

Economic power also often translates into political power. Corporate lobbying shapes the political system in a range of ways, from agenda control

178. *See generally* GANESH SITARAMAN, THE CRISIS OF THE MIDDLE-CLASS CONSTITUTION: WHY 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 generally Ganesh Sitaraman, *The Puzzling Absence of Economic Power in Constitutional Theory*, 101 CORNELL L. REV. 1445 (2016).

179. *See generally* Sonia K. Katyal, *Democracy & Distrust in an Era of Artificial Intelligence*, 151 DAEDALUS 322 (2022) (arguing that "AI decision-making poses a . . . challenge to democracy's basic framework").

180. 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. 932, 956-58 (1919); *see also* James Weinstein, *Participatory Democracy as the Central Value of American Free Speech Doctrine*, 97 VA. L. REV. 491 (2011) (arguing "that contemporary American free speech doctrine is best explained as assuring the opportunity for individuals to participate in the speech by which we govern ourselves," and describing this account as "both descriptively powerful and normatively attractive").

to substantively forestalling regulation.[181] 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.[182] Political scientists have found that companies and trade associations have an outsized influence on politics.[183] These are concerns across areas of policy, including in AI.[184]

As noted above, economic power at the individual level is also a form of political power. A voluminous literature in political science finds 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.[185] 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.[186] 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, self-preferencing, and discrimination, as well as questions

181. *See generally* LEE DRUTMAN, THE BUSINESS OF AMERICA IS LOBBYING: HOW CORPORATIONS BECAME POLITICIZED AND POLITICS BECAME MORE CORPORATE (2015) (describing lobbying and its impact).

182. *See generally* ALYSSA KATZ, THE INFLUENCE MACHINE: THE U.S. CHAMBER OF COMMERCE AND THE CORPORATE CAPTURE OF AMERICAN LIFE (2015) (describing how the Chamber, whose members include large U.S corporations, influences politics and policy).

183. KAY LEHMAN SCHLOZMAN, SIDNEY VERBA & HENRY E. BRADY, THE UNHEAVENLY CHORUS: UNEQUAL POLITICAL VOICE AND THE BROKEN PROMISE OF AMERICAN DEMOCRACY 404-11 (2012).

184. *See*, *e.g.*, Yochai Benkler, *Don't Let Industry Write the Rules for AI*, 569 NATURE 161, 161 (2019).

185. SCHLOZMAN, VERBA & BRADY, *supra* note 183, at 13-21, 117-33.

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

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 the future of democracy.

III. LESSONS FOR GOVERNANCE

Our account of the industrial organization of AI and the drawbacks of an AI oligopoly yields five important lessons. First, 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, 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. 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 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. And finally, open-source AI foundation models are unlikely to fully address the problems of an unregulated AI oligopoly.

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 are both easily discerned and a function of traits inherent to the technology and its industrial environment. 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."[187] This is sometimes referred to as the "pacing problem"—the idea that the pace of innovation is beyond the capacity of regulators.[188]

We disagree.[189] 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—for example, by declining to impose a separations rule between those layers at this time—while still protecting against other likely, foreseeable harms of concentration in these layers through, say, interoperability rules, nondiscrimination rules, 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.[190] 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 allows the industry to develop norms and standards that are guided by its collective private

187. Tom Wheeler, *The Three Challenges of AI Regulation*, BROOKINGS (June 15, 2023), https://www.brookings.edu/articles/the-three-challenges-of-ai-regulation [https://perma.cc/FE5R-NE3E].

188. 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 generally Maria Nordström, *AI Under Great Uncertainty: Implications and Decision Strategies for Public Policy*, 37 AI & SOC'Y 1703 (2022).

189. For another recent and thorough rebuttal to this argument, see Anu Bradford, *The False Choice Between Digital Regulation and Innovation*, 119 NW. UNIV. L. REV. 377 (2024).

190. DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY 19 (1980). For an application to AI, see Guihot, Matthew & Suzor, *supra* note 188, at 422.

interests, which may not align with the public interest.[191] Indeed, as Wheeler concludes, there were significant externalities of the private incentives that governed conduct in online markets: "market concentration," "invasion[s] of personal privacy," "user manipulation, and the dissemination of hate, lies, and misinformation."[192] Without regulation, we may well see similar effects in AI contexts.

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 are essential tools and important complements to ex post antitrust enforcement on a case-by-case basis.

Some have argued that we should embrace an approach of "permissionless innovation,"[193] allowing these companies to run amok until they cause substantial harm—and only then seeking to redress harms 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 backward at harms that had taken place, helped shape competitive markets by deterring anticompetitive behavior.[194] At the same time, antitrust enforcement under our prevailing legal standards is not likely to be ideal. 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

191. *See* Wheeler, *supra* note 187.

192. *Id*.

193. *See, e.g.*, Adam Thierer, *Getting AI Innovation Culture Right*, R ST. INST. 1-2 (Mar. 30, 2023), https://www.rstreet.org/wp-content/uploads/2023/03/Final-Study-No.-281.pdf [https://perma.cc/QWW2-GWJS] (arguing for a permissionless innovation approach).

194. Tim Wu, Opinion, *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 [https://perma.cc/RBW9-H9XP].

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.[195] 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,[196] and has required plaintiffs to show that the defendant could likely recoup its losses—even in a case with clear evidence of predatory pricing.[197] This judicial skepticism may make predation cases less likely to be successful, at least until it is far too late.

Another pricing doctrine focuses on "price squeezes," which occur when a vertically integrated firm with market power in an upstream business line charges high prices to downstream competitors. Power in the upstream market allows the firm to benefit from its own vertically integrated downstream business—while raising costs for competitors.[198] Supply squeezes are similar but involve the refusal to sell or to prioritize the sale of goods during a time of shortage.[199] In the AI context, vertical integration or partnerships across the technology stack raise the possibility of anticompetitive squeezes. Apple's partnerships with TSMC have raised questions about the semiconductor manufacturer's preferencing of Apple over other chip consumers.[200] 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

195. RICKS ET AL., *supra* note 42, at 226.

196. Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574, 589 (1986) ("[T]here is a consensus among commentators that predatory pricing schemes are rarely tried, and even more rarely successful.").

197. *See* Brooke Grp. Ltd. v. Brown & Williamson Tobacco Corp., 509 U.S. 209 (1993).

198. HERBERT HOVENKAMP, PRINCIPLES OF ANTITRUST 293-94 (2017).

199. *Id*.

200. *See supra* text accompanying notes 135-136.

subscriber line ("DSL") and internet service provider ("ISP") businesses.[201] 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.[202] This made it impossible for these would-be rivals to compete effectively. 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.[203]

Related to the refusal to deal is the essential-facilities doctrine. Under that doctrine, a firm that controls an "essential facility" must give reasonable access to users, even if those users are the firm's competitors. The doctrine requires (1) that the essential facility be controlled by a monopolist, (2) that it be infeasible for a competitor to replicate the facility, (3) that the competitor be denied access to the facility, and (4) that it be feasible for the monopolist to provide such access.[204] 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. Such barriers give firms in these areas considerable power—and with it, 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,[205] it has remained disfavored by leading antitrust experts.[206] The Supreme Court has also been skeptical of the doctrine. In *Verizon Communications Inc. v. Law Offices of Curtis V. Trinko, LLP*,[207] the Court observed that "[c]ompelling [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

201. Pac. Bell Tel. Co. v. LinkLine Commc'ns., Inc., 555 U.S. 438, 442-44 (2009).

202. *Id*. at 443-44.

203. *Id*. at 450-51.

204. MCI Commc'ns Corp. v. Am. Tel. & Tel. Co., 708 F.2d 1081, 1132-33 (7th Cir. 1983).

205. *See* Nikolas Guggenberger, *Essential Platforms*, 24 STAN. TECH. L. REV. 237, 305-43 (2021).

206. Phillip Areeda, *Essential Facilities: An Epithet in Need of Limiting Principles*, 58 ANTITRUST L.J. 841, *passim* (1990); HOVENKAMP, PRINCIPLES, *supra* note 198, at 297 (noting that the essential-facilities doctrine is "one of the most troublesome, incoherent and unmanageable" pathways to antitrust liability).

207. 540 U.S. 398 (2004).

incentive for the monopolist, the rival, or both to invest in those economically beneficial facilities."[208] While the Court did not completely reject the essential-facilities doctrine, it also did not adopt it.[209]

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, indeed, 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 must 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.

In antitrust, a great deal of decisional power also 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

208. *Id*. at 407-08.

209. *Id*. at 410-11 ("We have never recognized [the essential-facilities] doctrine, and we find no need either to recognize it or to repudiate it here." (citations omitted)).

set of facts and perspectives when crafting rules to drive firm behavior and can design rules for different situations.[210] And while courts are (by design) insulated from political accountability, 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 of 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 technology stack are, as noted, structurally inclined towards consolidation and concentration, meaning that underenforcement 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 regulation can harm innovation.[211] Our analysis of the AI technology stack

210. 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: A STRUCTURAL APPROACH TO REVERSING THE SECOND AGE OF MONOPOLY POWER 1 (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 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").

211. 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 193; 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 [https://perma.cc/P43T-T9N9] (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

not only undermines 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 services for AI) can leverage that power 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 by restricting the diversity of the downstream product market—first, by pushing competitors out, and then second, by chilling investment and entry into the downstream market. Indeed, economists have modeled how technology platforms have created "kill zones" in which venture capitalists decline to fund startups in markets where the dominant platform is likely to "crush" any new entrant.[212]

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.[213] 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, 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.

CONSTRUCTIONS OF INFORMATIONAL CAPITALISM 90-92 (2019) (summarizing—and ultimately dismissing—such arguments).

212. Kamepalli , Rajan & Zingales, *supra* note 156.

213. Tejas N. Narechania & Tim Wu, *Sender Side Transmission Rules*, 66 FED. COMMC'N L.J. 467, 470-74 (2014).

D. The Importance of Governing Market Structure

Our analysis of the AI technology stack also 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. And to the extent one believes 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 a focus on application conduct. This is for two reasons. First, understanding the structure of AI's technology stack allows us to identify locations in the stack where regulatory interventions can be most helpful for addressing downstream conduct issues. For example, if a concern stems from the quality of data that goes into training a model,[214] 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, holding cloud providers or model providers liable for downstream uses could force them to develop systems to screen potential users; requiring licensing of all users at the cloud or model layer could restrict users to those with training.[215]

Second, in some cases, confronting issues of market structure might help address conduct harms at the application layer by increasing the diversity of options at the application layer or regulating firms to ensure quality at the application layer.[216] Consider 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 reduce investments in improving accuracy because it has no reasonable

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

215. 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.

216. *See* Narechania, *supra* note 23, at 1592-95 (2022) ("[M]any of the now well-known problems attending to machine learning-based applications might be understood as problems of market power.").

competitors.[217] 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 needed to develop a workable model.

In other words, where an application will tend to consolidate its control over an application market (e.g., due to vertical integration or spectacularly high data and compute costs) the application provider is not only likely to tend towards higher user prices, but also gains a respite from competitive pressures that typically force quality improvements. In such cases, either competition or regulation is necessary. A more competitive market—even at these lower layers in the stack—might spur improvements in application accuracy or greater privacy protections,[218] and also help restrain price increases. Regulations, designed to ensure reasonable prices or spur innovation and quality improvements, might alternatively help prevent application harms. We should be clear about the scope of our claim: We do not mean to suggest competition or regulation is sure to address concerns about conduct harms like algorithmic bias and discrimination. But market structure interventions may be critically important complements to regulations that seek to directly address concerns about bias, discrimination, and privacy.[219]

E. The False Promise of Open-Source AI Competition

Some distinguished scholars and a number of AI-sector firms have argued that open-source AI models will help to "comba[t] market

217. *Cf.* Solow-Niederman, *supra* note 188, at 641 & n.33.

218. *See, e.g.,* Yifei Wang, Competition and Privacy 2-3 (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4766344 [https://perma.cc/UET5-FZQQ] (demonstrating a causal relationship between competition and privacy).

219. *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 on society"). We note that there is a voluminous literature on such matters, of which we have only begun to scratch the surface in the sources cited in *supra* notes 1-20, among other sources cited throughout this Article and elsewhere.

concentration," "promot[e] competition," and "distribut[e] power."[220] We think such claims go too far.[221] Policymakers should not conclude that open-source AI will completely address the problems with an unregulated AI oligopoly for four reasons.

First, "open source" is not a binary between open and closed, but rather a spectrum. AI's development so far has led to "gradients of openness."[222] Models have been released along the full spectrum from fully closed to fully open. Without specificity into what exactly is open, claims that "openness" will lead to competition are overbroad. Indeed, some open-source resources expressly include limits ensuring they will not be used to challenge their proprietors.[223]

Second, our analysis shows that the AI technology stack is concentrated beneath the model layer, in the hardware and cloud layers. The existence of open-source models does not address concentration in these other, lower layers.

Third, and relatedly, without a significant change in the industrial organization of those layers, any open-source AI foundation model will be dependent on an oligopoly of the biggest technology companies. If big

220. *See* Rishi Bommasani et al., *Considerations for Governing Open Foundation Models*, STAN. UNIV. HUM.-CENTERED A.I. 1-3 (Dec. 2023), https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf [https://perma.cc/9JA2-ZV4X] (making these points). Creative Commons, EleutherAI, GitHub, Hugging Face, LAION, and Open_Future argue that "open source development can enable competition and innovation by new entrants and smaller players." *See* Creative Commons, Eleuther AI, GitHub, Hugging Face, Laion & Open Future, *Supporting Open Source and Open Science in the EU AI Act*, CREATIVE COMMONS 4 (July 2023), https://creativecommons.org/wp-content/uploads/2023/07/SupportingOpenSourceAndOpenScienceInTheEUAIAct.pdf [https://perma.cc/K2AE-KSAE].

221. For other skeptical accounts, see Derek Slater & Betsy Masiello, *Will Open Source AI Shift Power from 'Big Tech'? It Depends.*, TECH POL'Y PRESS (June 16, 2023), https://www.techpolicy.press/will-open-source-ai-shift-power-from-big-tech-it-depends [https://perma.cc/MRH3-8XVM]; and Widder et al., *supra* note 122.

222. Irene Solaiman, The Gradient of Generative AI Release: Methods and Considerations 4-6 (Feb. 5, 2023) (unpublished manuscript), https://arxiv.org/pdf/2302.04844 [https://perma.cc/LF72-HZ3A] (describing the levels of access to generative AI models); *see also* Bommasani et al., *supra* note 220, at 3 (describing the "gradient" of AI model releases).

223. *See LLaMA 3.1 Community License Agreement*, *supra* note 121.

technology companies remain unregulated in the hardware and cloud layers, they would have the power to put firms based on open-source software out of business. As we have illustrated above, tactics like self-preferencing, tying of products, and integration of their own vertically integrated models into their applications could leave open-source models and the applications built upon them with smaller markets and prevent them from gaining a significant share—or even a foothold—in the marketplace. There is already evidence of these dynamics in related areas. Amazon, for example, released DocumentDB in 2019, an AWS database service based on an open-source database called MongoDB. In essence, Amazon was able to appropriate the efforts of open-source developers to develop its own competitor.[224] Indeed, many of the most notable open-source products have failed to compete effectively with big, integrated technology companies. Linux has not challenged Microsoft Windows or Apple iOS as an operating system at scale.[225] Web browser Firefox may have spurred innovation, but its features have been integrated into dominant browsers Google Chrome, Apple Safari, and Microsoft Edge; Firefox now trails all of these providers in browser market share.[226] Given that the proprietary models are integrated with business lines up and down the AI technology stack, open-source models may quickly fall behind these vertically integrated proprietary models,

Fourth, open-source development can itself be a strategy for maintaining their dominance. Google's Android operating system offers one example. Because Android OS is open source, Google benefits from platform network effects, as developers create many different applications for the operating system, helping Google compete with Apple's own offerings.[227] But that does not mean that the mobile operating systems markets are competitive. Rather, Google Android and Apple iOS are currently the only viable options in the market. In the AI context, developers rely on software frameworks to engage with data and develop models. PyTorch and TensorFlow, developed by Meta and Google, respectively, dominate in this

224. Widder et al., *supra* note 122, at 13.

225. Scharon Harding, *Linux Market Share Passes 4% for First Time; MacOS Dominance Declines*, ARS TECHNICA (Mar. 5, 2024), https://arstechnica.com/gadgets/2024/03/linux-continues-growing-market-share-reaches-4-of-desktops [https://perma.cc/3NSA-QAFB].

226. *Browser Marketshare Worldwide*, STATCOUNTER GLOBALSTATS, https://gs.statcounter.com/browser-market-share [https://perma.cc/F259-9GDM].

227. Widder et al., *supra* note 122, at 13.

arena, with PyTorch the leader among academic researchers.[228] Because Meta released PyTorch publicly, it has gained widespread adoption.[229] Meta ultimately benefits from developers creating AI systems and innovations built on its framework because Meta can "more easily integrate and commercialize academic AI models, and others developed, tuned, or deployed using PyTorch."[230] In short, open source can be used to disrupt dominant platforms—but it can also be used by dominant firms to extend their platforms and entrench their status. In such contexts, open source is not an opponent of big technology companies, but their handmaiden.

In light of these dynamics, we should not expect open-source models to achieve the antimonopoly benefits that commentators have suggested absent changes to the industrial structure of the AI stack.

*

Our analysis of AI's industrial organization—the structure of the firms and markets that comprise AI's supply chain—yields several key insights: (1) significant aspects of the AI supply chain are sufficiently stable to warrant regulatory intervention, ongoing technical developments notwithstanding; (2) ex ante regulatory mechanisms are likely to be more effective than ex post enforcement; (3) regulatory frameworks have the potential to foster greater downstream innovation, as the current market structure may be inhibiting such activity; (4) concerns surrounding AI, including bias and privacy, may, in part, stem from market-structure concerns; and (5) open-source development, while valuable, may serve as an imperfect substitute for other forms of competition due to AI's industrial organization. Given these conclusions, we turn next to possible policy approaches.

IV. AN ANTIMONOPOLY APPROACH FOR ARTIFICIAL INTELLIGENCE

As we have seen, the AI technology stack is characterized by concentration in many of its layers. This concentration seems likely to persist. 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. Ex ante tools can also spur innovation and help address conduct harms from AI applications. In this final Part, we outline the antimonopoly tools—industrial policy, NPU

228. *Id*. at 6.

229. *Id*. at 7.

230. *Id*.

rules, public options, and cooperative governance—that can help govern the AI 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 within the United States. The bipartisan Chips and Science Act of 2022[231] 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.[232] The Commerce Department's Chips for America program seeks to use federal funds to encourage private sector investment in order to develop at least two large-scale clusters for fabrication of chips.[233]

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

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

232. *Id*.

233. CHIPS FOR AMERICA, NAT'L INST. STANDARDS & TECH., VISION FOR SUCCESS: COMMERCIAL FABRICATION FACILITIES 1, (Feb. 28, 2023), https://www.nist.gov/system/files/documents/2023/02/28/Vision_for_Success-Commercial_Fabrication_Facilities.pdf [https://perma.cc/TGB8-V9G9].

Whereas the Chips and Science Act is aimed at spurring domestic development to address concentration in production capacity, other efforts are aimed more directly at national security concerns. For example, the Biden Administration has restricted sharing advanced semiconductor technologies with certain Chinese entities. Michael Schuman, *Why Biden's Block on Chips to China Is a Big Deal*, ATLANTIC (Oct. 25, 2022), https://www.theatlantic.com/international/archive/2022/10/biden-export-control-microchips-china/671848 [https://perma.cc/UJ24-PZSY]. And the Administration is also reportedly preparing restrictions on outbound investment to Chinese technology firms. Jack Stone Truitt, *Biden Executive Order on Investments in China Faces Hurdles*, NIKKEI ASIA (June 10, 2023), https://asia.nikkei.com/Politics/International-relations/US-China-tensions/Biden-executive-order-on-investments-in-China-faces-hurdles [https://perma.cc/YQ35-4CS3].

further extend their lead. But industrial policies could target new, smaller, and more innovative actors, in which case these policies can facilitate competition rather than entrench market power.[234] It is too early to tell 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 Science Act—could consider market diversification and competition as a critical element in evaluating candidates for federal grants.[235] There is also reason to believe that the federal government will take an antimonopoly approach in its industrial policy efforts. The Office of Management and Budget's draft memorandum to federal agencies on AI advises agencies to ensure that procurement "promote[s] opportunities for competition" and doesn't "improperly entrench incumbents."[236] The draft memorandum provides few details but primes agencies to promote, rather than hinder, competition in their contracting and procurement practices, including by requiring compliance with rules steeped in the tradition of regulating networks, platforms, and utilities.[237]

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.[238] NPU regulations provide a legal framework that can help build NPUs at scale, ensure continuity of

234. *See* Philippe Aghion et al., *Industrial Policy and Competition*, 7 AM. ECON. J.: MACROECONOMICS 1 (2015).

235. 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, *supra* note 233, at 1, 5.

236. OFF. OF MGM'T AND BUDGET, PROPOSED MEMORANDUM FOR HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES: ADVANCING GOVERNANCE, INNOVATION, AND RISK MANAGEMENT FOR AGENCY USE OF ARTIFICIAL INTELLIGENCE 21 (Nov. 2, 2023).

237. *Id*. at 21-22 (advising agencies to promote interoperability, prevent self-preferencing, avoid vendor lock-in, and prevent the use of data to entrench dominance); *see also* Deirdre K. Mulligan & Kenneth A. Bamberger, *Procurement as Policy: Administrative Process for Machine Learning*, 34 BERKELEY TECH. L.J. 773, 778-80, 785-98 (2019) (discussing how government procurement processes focus on promoting management goals like competition).

238. RICKS ET AL., *supra* note 42, at 7-10.

service, prevent monopolistic and oligopolistic abuses, avoid destructive competition, ensure widespread access, promote commercial development, and sustain democracy.[239] These regulations operate primarily ex ante, that is, by structuring the market (often to favor greater downstream competition), identifying likely, foreseeable harms, and establishing rules to prevent those harms *before* they arise.[240] In this subsection, we describe how selected NPU tools could be helpful in addressing the downsides of an AI oligopoly.

1. Structural Separations.

Structural separations "limit the lines of business in which a firm can engage."[241] The central benefit of structural separations is preventing a business from self-preferencing or leveraging its 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.[242] 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 profit leveraging across business lines, structural separations also limit the concentration of economic power and promote a diverse business ecosystem of users of the platform.[243] Perhaps most importantly, structural separations are more administrable than many other policies. 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.

239. *Id.* at 11-21.

240. More precisely, these rules are designed to operate in contexts where policymakers determine that the benefits of the regulatory approach outweigh their costs, accounting for the possibilities that ex post enforcement will be too little, too late, as well as for the possibility that a regulatory approach may miss its mark (by, say, creating regulatory barriers to entry that undermine other entrants). That is, accounting for the benefits and costs, including error costs, an ex ante approach is superior.

241. *Id.* at 28.

242. Hepburn Act of 1906, Pub. L. No. 59-337, 34 Stat. 553 (1906) (with one commodity exception).

243. 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).

With respect to AI, there are a number of places where structural separations could be useful.[244] 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 for AI) that is open for all kinds of uses—like electricity—and ensure that those providers cannot prioritize their own downstream business lines over those of their competitors. Separation could also spur cloud providers to innovate on their cloud offerings, rather than rely on innovation from vertical integration.[245] This would, in turn, also facilitate innovation in downstream markets where cloud users could develop a range of products and services rather than be pushed into the cloud company's system.[246]

2. Nondiscrimination, Open Access, and Rate Regulation.

One alternative to structural separation requirements is nondiscrimination and equal access rules, which are sometimes coupled with rate regulation.[247] 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

244. *Cf.* William P. Rogerson & Howard Shelanski, *Antitrust Enforcement, Regulation, and Digital Platforms*, 168 U. PA. L. REV. 1911, 1934-36 (2020) (discussing how line-of-business restrictions to digital platforms can break up the potential for monopolies without risking significant harms to competition).

245. If regulators decide that integrating chips and cloud is desirable for effective service provision, then separating chips/cloud from higher layers in the stack would encourage innovation across both layers together while preserving the innovative potential of competition further up the stack. *Cf.* Wu, *supra* note 152, at 141-42 (explaining that vertical integration between ISPs and facilities ownership might be desirable, and so advocating for a nondiscrimination regime to protect downstream innovation).

246. *See* Slee, *supra* note 142, at 122 ("In some industries the essential infrastructure is heavily regulated and controlled, while services built on that infrastructure are opened for innovation . . . . Core banking functions are strictly regulated . . . while many countries are experimenting with open banking laws to permit innovation on top of this infrastructure.").

247. RICKS ET AL., *supra* note 42, at 24-26.

business lines.[248] 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.[249] All users must also be treated similarly in terms of price.[250] Historically, nondiscriminatory pricing rules required firms to file their prices, called "tariffs," and make them publicly available.[251] Price transparency 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 otherwise charge prohibitive prices to undermine nondiscriminatory access requirements.[252] 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.[253]

Nondiscrimination and equal-access rules complement structural separations when a business has market dominance or acts as a platform for downstream activity. The reason is that a structurally separated platform can still pick and choose its users or charge differential 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 second-best strategies for addressing self-preferencing concerns because of administrability issues. While neutrality between business lines should prevent self-preferencing, in practice, it is more difficult for regulators to police and enforce nondiscrimination rules than structural separation requirements.[254] The latter requires regulators to monitor or audit specific business practices and identify violations of pricing or treatment rules—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 prophylactic: They prevent any commingling of business lines and thus can be more easily administered.

248. *Id*. at 24, 26, 29.

249. Ganesh Sitaraman, *Deplatforming*, 133 YALE L.J. 497 (2023).

250. RICKS ET AL., *supra* note 42, at 24.

251. *Id*.

252. *Id*. at 24-25.

253. *Id*. at 25.

254. *See, e.g.*, Rory Van Loo, *In Defense of Breakups: Administering a "Radical" Remedy*, 105 CORNELL L. REV. 1955, 2006-08 (2020) (arguing that "access remedies" are not necessarily easier to administer than corporate breakups).

In the AI context, nondiscrimination and equal-access rules could be adopted at multiple places in the stack. At the hardware layer, given the scarcity of chips, fabricators and designers could be required to serve customers equally, at least until chip fabrication becomes more available. At the cloud layer, cloud providers could 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 data lakes could also be subject to nondiscrimination and equal-access rules. This would enable more model developers to build and train new models. Foundation models and APIs could also be subject to such rules, allowing app developers to tweak those models to develop new products and services.

3. Interoperability Rules.

Interoperability rules lower barriers to entry and can thus stimulate competition by "allowing new competitors to share in existing investments" and "imposing sharing requirements on market participants."[255] In the telecommunications industry, 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 each other, no one provider could wield its network effects as a sword to effectively consolidate control over the entire market. Customers could choose any provider and still reap the benefits of a network that spanned the entire market. Rules that required 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 familiar lock-in effect: It is quite cumbersome to let all your contacts know you have a new phone number.[256]

Such requirements could be applied to AI. Recall that among 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

255. Narechania, *supra* note 23, at 1555.

256. NUECHTERLEIN & WEISER, *supra* note 150, at 57 (discussing number portability).

across many participating clients that keep their training data locally."[257] Rules that require a federated learning approach among competitors may be attractive to policymakers seeking to induce competition while ensuring that no single application, vertically integrated with the underlying model, uniquely benefits from improvements made through continuous or reinforcement learning.[258] 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 de-duplication,[259] not only to enable federated learning, but also to improve outcomes and protect user privacy.[260] Such forms of AI development may help to undermine the consolidation-driving network effects of the data sublayer.

Policymakers might also consider rules that improve interoperability among cloud platforms, easing transitions from one provider's system to another.[261] The lack of interoperability and the problems of switching are

257. *See TensorFlow Federated: Machine Learning on Decentralized Data*, TENSORFLOW, https://www.tensorflow.org/federated [https://perma.cc/H8AZ-G6P2] (describing one approach to federated learning).

258. This approach is distinct from the one adopted in Europe via Gaia-X, which predominantly regards federated data storage, for the purposes of complying with data localization requirements (e.g., rules that certain personal data be housed in certain locales). *See About Gaia-X*, GAIA-X https://gaia-x.eu/what-is-gaia-x/about-gaia-x [https://perma.cc/A9MN-HCUS] (describing Gaia-X as a "A Federated and Secure Data Infrastructure"). 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.

259. De-duplication refers to the process of cataloging and removing duplicate data entries. *See* Brandon J. Trout, *Infringers or Innovators? Examining Copyright Liability for Cloud-Based Music Locker Services*, 14 VAND. J. ENT. & TECH. L. 729, 733 (2012) (explaining that data de-duplication conserves "valuable storage space, bandwidth, and costs").

260. 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.

261. *See, e.g.*, *Cloud Services Market Investigation Qualitative Customer Research: Final Report*, COMPETITION & MKTS. AUTH. 44 (May 2024), https://assets.publishing.service.gov.uk/media/664f02634f29e1d07fadcd56/Cloud_Services_Market_Investigation_Qualitative_Customer_Research_Final_Report_.pdf [https://perma.cc/5HMR-HK75] (identifying technical barriers as one primary impediment to switching among cloud providers).

real concerns. Technologists, for example, have proposed entire systems, such as "Sky Computing," aimed at addressing the switching and interoperability costs associated with using different cloud providers.[262] 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 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 one from some new competitor). A common API, used across providers, could lower that developer's switching costs and thereby yield greater competition.[263] In all, interoperability among distinct providers can facilitate entry and foster innovation, giving rise to better outcomes for participants in the downstream model and application layers—and ultimately for consumers.

C. Public Options

Another policy tool for increasing competition and service reliability is a public option. Public options are publicly provided goods or services that coexist with private-market options, offered at some (often regulatorily) set price.[264] Public options can help ensure competition, as the public option disciplines private monopolists or oligopolists that might increase prices or reduce service quality.[265] Competition from private parties, in return, ensures that the public option provides high-quality service as well.[266] A

262. *See* Ion Stoica & Scott Shenker, *From Cloud Computing to Sky Computing*, 2021 WORKSHOP ON HOT TOPICS IN OPERATING SYS. 26, 27.

263. *Cf.* Google LLC v. Oracle Am., Inc., 593 U.S. 1, 34 (2021) ("Given the costs and difficulties of producing alternative APIs with similar appeal to programmers, allowing [copyright] enforcement [against Google] would make of the Sun Java API's declaring code a lock limiting the future creativity of new programs.").

264. SITARAMAN & ALSTOTT, *supra* note 39, at 27.

265. *Id*. at 38-40.

266. *See* E.S. Savas, *An Empirical Study of Competition in Municipal Service Delivery*, 37 PUB. ADMIN. REV. 717, 723 (1977) (finding that Minneapolis's waste collection service became more cost-efficient and productive after

public option also adds to the diversity of the sources of production (even if only slightly) thereby strengthening supply chain resilience and reliability.

In the AI context, a public option for cloud infrastructure could serve as a helpful complement or alternative to structural separations or nondiscrimination and equal access rules.[267] Because of high capital costs, network effects, and concerns from vertical integration, a public option for the 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 marginally increase competition, by offering an alternative to high-priced oligopoly providers. And it can help ensure that compute is affordable for 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.[268]

Notably, the National Science Foundation's ("NSF") proposal to offer a National AI Research Resource (NAIRR) has focused on public access to AI research. NAIRR would "democratize access to AI resources" and therefore "must primarily be sustained through Federal investment."[269] However, NSF's proposal is unclear on the degree to which NAIRR would be a public option, or whether the federal government would contract with private companies for critical AI services.[270] It suggests NAIRR provide a mix of

introduction of direct private waste-collection competition); *see also* E.S. Savas, *Intracity Competition Between Public and Private Service Delivery*, 41 PUB. ADMIN. REV. 46, 48 (1981) (arguing that contracting with private waste collections can offer a city "a yardstick against which to measure the performance of its own municipal agency").

267. *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY 208-09 (2015) (arguing that public alternatives to private-sector algorithms could undermine search monopolies and lead to more transparent market actors).

268. Nikkei Staff Writers, *Japan's METI to Build New Supercomputer to Help Develop AI at Home*, NIKKEI ASIA (July 24, 2023), https://asia.nikkei.com/Business/Technology/Japan-s-METI-to-build-new-supercomputer-to-help-develop-AI-at-home [https://perma.cc/2DFU-MZBZ].

269. NAT'L A.I. RSCH. RES. 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), https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf [https://perma.cc/6ZLC-RMF5].

270. *See* AI Now Inst. & Data & Soc'y Rsch. Inst., *Democratize AI? How the Proposed National AI Research Resource Falls Short*, AI NOW (Oct. 5, 2021),

computational resources, including "commercial cloud" as an option.[271] 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.[272] Recently introduced legislation would create a NAIRR that offers "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 APIs.[273] This structure may require some amount of nonoligopoly 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 it is a true public option, rather than a government contract for researchers to purchase compute and other resources from oligopolistic cloud providers while further entrenching them.

The NAIRR legislation also includes provisions for data access,[274] and the federal government is considering several other initiatives aimed at releasing public datasets to support model development.[275] 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" without relying on their data.[276]

---

https://ainowinstitute.org/publication/democratize-ai-how-the-proposed-national-ai-research-resource-falls-short [https://perma.cc/P6U8-LNWB].

271. NAT'L A.I. RSCH. RES. TASK FORCE, supra note 269, at 31.

272. *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.").

273. CREATE AI Act of 2023, H.R. 5077, 118th Cong. § 5603(b); *see* Press Release, 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 [https://perma.cc/EFU9-KBPR].

274. *Id*.

275. *See, e.g.*, AI and Open Government Data Assets Request for Information, 89 Fed. Reg. 27411 (Apr. 17, 2024); *National AI Initiative Office Launches AI Researchers Portal*, HCP WIRE (Jan. 6, 2022), https://www.hpcwire.com/off-the-wire/national-ai-initiative-office-launches-ai-researchers-portal [https://perma.cc/9G3A-6MUY].

276. Gansky, Martin & Sitaraman, *supra* note 40.

D. Cooperative Governance

Cooperatives are firms that are owned by consumers, workers, or producers and that emphasize participatory membership, democratic control, and economic participation, among other core principles.[277] 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.[278] Today, many familiar U.S. companies are cooperatives, such as outdoor retailer REI, Sun-Maid Raisins, Land O'Lakes, State Farm Insurance, and Ace Hardware.[279]

One of the features of cooperatives is their potential to subvert monopoly power.[280] In the early nineteenth century, as Henry Hansmann and Mariana Pargendler have shown, corporations in NPU industries—turnpikes, canals, railroads, banks—were legislatively chartered monopolies.[281] This structure gave rise to standard monopoly concerns, such as monopoly pricing. Legislators responded by imposing restrictive corporate voting rights that placed power in the hands of consumer-owners.[282] 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, and directly addressed common concerns about monopoly pricing and services.[283] In the late nineteenth century, as capital became more available, general incorporation laws became widespread, and corporations grew to national scale, the

277. *Understanding the Seven Cooperative Principles*, NAT'L RURAL ELEC. COOP. ASS'N (Dec. 1, 2016), https://www.electric.coop/seven-cooperative-principles [https://perma.cc/4JEG-UP67] (also listing as principles of cooperatives: autonomy and independence; education, training, and information; cooperation among cooperatives; and concern for community).

278. *See* Bruce J. Reynolds, *Comparing Cooperative Principles of the U.S. Department of Agriculture and the International Cooperative Alliance*, U.S. DEP'T OF AGRIC. 2 (June 2014), https://www.rd.usda.gov/files/BCP_RR231.pdf [https://perma.cc/4RM9-V2ES].

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

280. *See* Sandeep Vaheesan & Nathan Schneider, *Cooperative Enterprise as an Antimonopoly Strategy*, 124 PENN. ST. L. REV. 1, 1, 23-25 (2019).

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

282. *Id*.

283. *See id*.

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.[284] Cooperative governance, Hansmann observes, acted as an alternative to "not only the costs of monopoly but also the costs of rate regulation."[285]

Cooperatives are an antimonopoly tool because they "accomplish vertical integration" in a way that limits exploitative conduct.[286] 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 to shareholders in the form of higher dividends or stock buybacks.[287] Cooperative governance shifts management's priorities 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 may distribute 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 is more consistent with the goals and values of its users.[288] 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, operate models, and share in the ownership of the cloud, all without fear that one of the big platforms will take these owners' ideas or raise prices for the utility services the platforms provide. One might even imagine a cooperative model where one earns stakes in a model or application by contributing data to its development—or alternatively, a cooperative-governance approach to a foundation model, in which users govern the model. These options for cooperatives in the cloud and model layers would help introduce

284. *See id*. at 945-55.

285. HENRY HANSMANN, THE OWNERSHIP OF ENTERPRISE 170 (1996).

286. Molk, *supra* note 279, at 912.

287. Vaheesan & Schneider, *supra* note 280, at 9.

288. *Cf*. Slee, *supra* note 142, at 122 (noting the successful and collaborative nature of Wikipedia and concluding that "something is working on Wikipedia that is not working at . . . Facebook[] or Amazon").

competition between the cooperative and private platforms, while simultaneously offering greater access to AI resources and distributing wealth more equitably.

*

In all, antimonopoly tools—such as industrial policy, NPU (networks, platforms, utilities) law, public options, and cooperative governance—can help bring greater competition to the layers of the AI stack. For instance, industrial policies addressing semiconductor scarcity can both highlight the importance of resilience and national security and can address the power of dominant firms. NPU law, with its longstanding tools—e.g., structural separation, nondiscrimination, and interoperability rules—offers a framework for regulating AI. Public options can complement these tools by promoting competition, setting a price floor, and fostering innovation. Finally, cooperative governance presents a more equitable way to manage AI-related enterprises, one that is more likely to reflect the values of its members.

CONCLUSION

Artificial intelligence has sparked considerable conversation and concern. Understanding the AI technology stack shows that 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 abuses of power, national security and resilience challenges, widening economic inequality, and political influence that can undermine democracy.

Antimonopoly tools 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."[289] Political leaders have allowed that approach to define technology in the early twenty-first century. The result has been a governance failure that has led to concentration and a range of economic,

289. Chris Velazco, Facebook *Can't Move Fast to Fix the Things it Broke*, ENGADGET (Apr. 12, 2018), https://www.engadget.com/2018-04-12-facebook-has-no-quick-solutions.html [https://perma.cc/TQY9-WUUJ]; *see generally* JONATHAN TAPLIN, MOVE FAST AND BREAK THINGS: HOW FACEBOOK, GOOGLE, AND AMAZON CORNERED CULTURE AND UNDERMINED DEMOCRACY, at epigraph (2017) (explaining this concept).

social, and political problems.[290] As policymakers debate governing AI early in its technological lifecycle, antimonopoly tools must be part of the conversation.

* * * * *

290. For examples of discussions of these problems (some of which are more rigorous than others), see generally ROGER MCNAMEE, ZUCKED: WAKING UP TO THE FACEBOOK CATASTROPHE (2019); ZUBOFF, *supra* note 116; and ROB REICH, MEHRAN SAHAMI & JEREMY M. WEINSTEIN, SYSTEM ERROR: WHERE BIG TECH WENT WRONG AND HOW WE CAN REBOOT (2021).