Model Monopolies: Assessing Natural Monopoly Dynamics and their Regulatory Implications in the Foundation Model Industry

George Colville

Word count: 14,995

Thesis submitted in partial fulfilment of the requirement for the degree of MSc in Social Science of the Internet at the Oxford Internet Institute at the University of Oxford

Trinity term, 2024

Abstract

This paper analyses the general characteristics of the foundation model industry from an industrial organisation perspective focused on natural monopoly. It identifies the presence of natural monopoly dynamics in foundation model production and examines the implication of these for competition policy. Recognising the dynamic and uncertain trajectory of the products proposed by the industry, it seeks to identify enduring dynamics likely to characterise it. Theoretical insights are used to examine whether (a) natural monopoly dynamics are likely to characterise the foundation model industry, generating non-temporary market power, and, based on this analysis, (b) whether any specific market regulation beyond general competition law is warranted. Synthesising over a century of natural monopoly theory and drawing extensively from the computer science literature, this paper argues that powerful natural monopoly dynamics characterise foundation model production. Unlike previous naturally monopolistic digital industries characterised by demand-side network effects, it is supply-side economies of scale that drive these dynamics in the foundation model context. Recognising that the nature of future foundation models is fundamentally uncertain when it comes to the question of product differentiation, a taxonomical treatment of different possible differentiation scenarios demonstrates how underlying supply-side natural monopoly dynamics nevertheless drive the industry towards concentration. The regulatory consequences of these findings are considered, with a preference shown to maximising contestability through the application of effective traditional ex-post competition law complemented by targeted ex-ante obligations.

Table of Contents

Introduction

"Big Tech Is Bad. Big A.I. Will Be Worse."[1]

"Big Tech's Budding AI Monopoly"[2]

"The looming AI monopolies"[3]

"The big tech firms want an AI monopoly"[4]

Historical precedent in digital markets combined with rapid growth in investment and public and regulatory interest in artificial intelligence (AI) has generated headlines and widespread concerns about concentrations of market power. However, most of these sensational headlines were written in the absence of rigorous scholarly analysis. They rely on analogies to fundamentally different digital technologies, markets, and industries and on the observation that Big Tech monopolists are active throughout the industry's value chain.

Theoretical contributions have a long tradition of informing competition policy and regulatory practice by providing a deeper understanding of market dynamics, systems, and causal relationships that empiricism alone cannot capture. This paper therefore conducts a theoretical analysis of competition concerns in the 'AI industry,' bringing insights from computer science literature to inform its economic and regulatory examination of the issue. Synthesising over a century of natural monopoly theory, it examines whether monopoly dynamics are inherent to foundation model production and considers the extent to which these are liable to cause market failures. It draws on these findings to reflect on what appropriate regulatory responses to these potential concerns might resemble, bearing in mind that dominance is legal and can be efficient, but abuse of dominance is not.

Using the concept of 'artificial intelligence' to characterise an industry, however, confounds rather than enables useful analysis of competition concerns from an economic perspective. Part I of this paper problematises that concept, in general and specifically for analytical

[1] New York Times (2023)
[2] Wall Street Journal (2024)
[3] POLITICO (2024)
[4] The Guardian (2024)

purposes, due to its broadness, its lack of an accepted or acceptable definition, and its ubiquitous misapplication. Moving away from the notion that there is a 'market for AI' or an 'AI industry', it focuses on foundation models, defined as models pre-trained on large datasets that can be fine-tuned and adapted to a wide range of downstream tasks (Bommasani & Liang, 2021).

This sub-field deserves particular attention not because it is guaranteed to be the most effective, widespread, or influential technical method for developing AI applications, though that possibility exists. It is instead of likely enduring relevance because it has been widely interpreted as such at a pivotal moment, driving powerfully influential investment and intense regulatory scrutiny at a time when many jurisdictions are drafting wide-reaching AI regulations (Stanford Institute for Human-Centered Artificial Intelligence (HAI), 2024). As AI applications start to be implemented more widely and regulators begin to address significant questions as the field moves away from its predominantly academic and research-based origins, foundation models dominate regulatory and political imaginaries and debates about AI, including within competition authorities (UK Competition and Markets Authority, 2023a). Ensuring these are appropriately framed is vital.

Many concerns voiced about the potential for competitive harm are based on largely impressionistic empirical observations about the dominant presence of a small group of firms across the foundation model value chain. While this is legitimate cause for concern, theoretical analysis of the industry's underlying dynamics has particular value given the pace of change in the industry. A deeper theoretical understanding of the forces in play enables the development of effective policy responses by illuminating the causes and consequences of concentration, not just its presence. Moving beyond static empiricism is especially necessary when seeking to understand industries whose natural tendencies are distorted by short-term anomalies including irrational behaviour and the presence of wealthy sponsors, making data-driven analyses challenging.

As outlined in Part II, examination of market concentration in the foundation model industry has, to date, been conducted primarily by legal and regulatory scholars, national competition agencies, and other actors writing from a policy perspective. Though they provide valuable insights and a number of valid findings, their work relies on relatively cursory economic analyses, mostly limited to listing the presence of market concentration risk factors including economies of scale and network effects. Likely due to previous developments and concerns in

platform markets (see e.g. Lehdonvirta, 2022), most focus on the vertical configuration of the industry, examining how concentration at different levels – notably in cloud computing – might have downstream effects on the industry as a whole. But truly effective analysis of concentration in the wider industry first requires an understanding of the dynamics at play at each level of its value chain.

Foundation model production, in many ways the industry's nucleus, has received insufficient attention in this respect. This paper addresses this oversight by conducting a theoretically grounded industrial organisation analysis of foundation model production, evaluating the extent to which its features give rise to natural monopoly dynamics. It is argued that the foundation model industry differs substantially from other digital monopoly markets due to lesser influence from network effects and more significant sunk fixed costs. Part III introduces the basics of the industrial organisation theory this analysis relies on. While systematically predicting and modelling market structure by reference to observables including technology is an impossible task, it certainly is true that the technological features of a product's means of production and the technological affordances of the product itself decisively shape that product's equilibrium market structure (Mosca, 2008; Sutton, 1998). Having established that network effects have a limited role to play in driving natural monopoly dynamics in Part IV.A, Part IV.B evaluates the cost structure of foundation model production, establishing that foundation model production exhibits strong static natural monopoly properties due to the presence of significant sunk costs and low marginal costs.

Part IV.C examines the consequences of these static natural monopoly properties in the absence of an incumbent monopolist. In a dynamic situation, the promise of natural monopoly rents often results an initial period of intense competition as firms compete for the market and potential monopoly rents. War of attrition scenarios are presented, demonstrating how price and innovation competition in the industry influence prices, welfare, and efficiency. The contestability of foundation model monopolies is then considered to determine the extent to which hypothetical monopoly positions are likely to translate to market failures in the foundation model industry given its characteristics.

To make a useful contribution and avoid excessive conjecture, academic humility and a recognition of the pervasive uncertainty surrounding the foundation model industry's future is essential. While the fundamental nature of the industry's cost structure is unlikely to change, the nature of the product(s) proposed is entirely unpredictable. The question of the future of

product differentiation looms large and unanswerable. Section IV.E addresses this by acknowledging the uncertainty from the outset, using it to frame its analysis by systematically considering the competition consequences of different specialisation outcomes. A taxonomical approach is adopted to illustrate how varying degrees and dimensions of product differentiation can alter competitive dynamics within the industry. Despite resulting variations in outcomes, particularly concerning scale, the industry is anticipated to remain naturally characterised by concentration.

The combination of an uncertain future landscape due to differentiation and innovation, but confidence in the cost-generated natural monopoly dynamics of foundation models makes proactive regulation of this area challenging. The preceding natural monopoly analysis raises two interlinked questions for regulatory attention. First, ensuring the competitive process for model monopolies or oligopolies is efficient is essential, both for ensuring the most efficient firms emerge as market leaders, and because it maximises the contestability of those markets even when incumbents are installed. I argue that revived and proactive ex-post enforcement of competition law can be highly effective, but that it may nevertheless benefit from supplementary assistance from regulation providing for expedited sector-specific ex-post enforcement and some ex-ante provisions which mandate behaviour in input and distribution industries, notably cloud computing. I argue that the European Union's Digital Markets Act (DMA) already provides for most of these. Second, the potential for ex-ante or utility-style regulation to correct competitive harms that follow from monopoly is considered. While price regulation is unfeasible, it is suggested that ex-ante regulation will be necessary to prohibit foundation model monopolists and oligopolists from leveraging their control over models with essential-facility-like positions to influence upstream competition in the cloud computing service industry and downstream competition in the fine-tuned model and application development industry.

I. AI and foundation models: a primer

Applying economic theory to the so-called 'AI industry' requires not only a knowledge of economics but also a clear technical and historical understanding of the technologies that fundamentally define the dynamics of competition within it. Too often, misconceptions about the nature of AI preclude otherwise estimable work from fulfilling its potential. In particular, the widespread tendency to treat AI and even foundation models monolithically, without recognising the diversity of the products and technologies that fall within these labels, impedes useful and precise analysis and thinking.

AI: a problematic analytical concept

In the midst of an 'AI summer' ushered in by successes in deep learning and the emergence of the first foundation model-based applications in the early 2020s, future imaginaries have run wild across industry, politics, and wider public discourse (Floridi, 2020; Kautz, 2022). Transformer-based large language models (LLMs) trained using deep learning on immense datasets and generative personal assistant applications built on them launched an 'AI boom', with massive increases in interest and investment (HAI, 2024). However, 'artificial intelligence' is a misleading term which lends itself to false conflations between machine intelligence and human-like understanding, reasoning and consciousness, especially outside technical circles (Fortuna & Gorbaniuk, 2022; Hirsch-Kreinsen, 2023). A powerful and evocative label, its ambiguity also makes it analytically problematic as anthropomorphic understandings of the technology mislead both its developers (Salles et al., 2020) and the wider public (Blut et al., 2021).

The field of AI research, and its industrial equivalent, consists of multiple sub-fields with extremely diverse goals, methods, technologies, levels of complexity and performance, and applications. It encompasses more than the LLM-dominated view that captured public imagination following the release of ChatGPT. For instance, rule-based expert systems have been used since the 1980s and remain valuable for certain applications (Bobrow et al., 1986). The AI umbrella also encompasses decision-making systems used in autonomous vehicles (Ma et al., 2020), convolutional neural networks used for image classification (Rawat & Wang, 2017), models which predict protein structures from amino acid sequences using deep

learning and attention mechanisms (Jumper et al., 2021), and reinforcement learning systems that can play games including Go at a superhuman level (Silver et al., 2016).

These research areas are labelled ‘AI’ for primarily historical reasons, not theoretical or technical ones. Many research breakthroughs in these applications aim to mimic or achieve human or superhuman intelligence (Wang, 2019). As products (or byproducts) of this endeavour, they inherit its title. AI is therefore best understood not as a single industry or technology, but as a broad field oriented around developing computing systems capable of performing tasks which otherwise require human intelligence.

A loosely identifiable AI-adjacent industry is growing rapidly in size and influence as technological breakthroughs lead to the development and marketing of products, and many players are simultaneously involved in multiple sub-fields (HAI, 2024). But satisfactorily defining an AI industry for analytical purposes, is challenging and unhelpful due to its broad and diverse nature. Many valuable contributions have suffered from attempts to generalise and analyse AI, often resulting in either mislabelling sub-fields or creating overly abstract analyses that are difficult to apply or use for policy development (see e.g. Gans, 2024; Marar, 2024; Narechania & Sitaraman, 2023). As Wang (2019) notes, the only real solution to this confusion is to find appropriate names for each of AI’s sub-fields and to start using them more rigorously rather than falling back on the AI hypernym.

Foundation models

This paper focuses on the fastest growing and most publicised major AI sub-field: foundation models. Implicitly referred to by many when they use the term ‘AI’, these are ‘models trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks’ through fine-tuning and integration into applications (Bommasani & Liang, 2021). This sub-field captured global techno-imaginaries following the public release of ChatGPT in late 2022, experiencing a nine-fold year-on-year increase in private investment in 2023. That year, foundation-model-related investment accounted for over a quarter of all AI-related private investment (HAI, 2024).

Foundation models are significant not only for their potential future impact but also for their perceived future impact, which directly influences investment and regulatory debate. They can be used as standalone systems or as building blocks for various purpose- and sector-

specific applications (Küspert et al., 2023). There are widespread expectations that they will be highly influential, widely used, and extremely profitable, especially if the most ambitious general-purpose models are widely implemented. Future imaginaries of foundation models have spurred massive investments, driven industries to adopt these systems, and led to claims that we are entering a 'foundation model era' (Soliani, 2024). These 'fictional expectations' (Beckert, 2016) influence economic and regulatory decisions, impacting investment, innovation, and regulatory development. Predictions that foundation models will be central to vast multi-actor economic networks are driving regulatory debate and activity, with regulators and legislators worldwide often reacting to developments and future imaginaries in this sub-field while drafting broader AI regulations.

General-purpose LLMs represent only one approach to developing pre-trained foundation models. While large players develop immensely powerful general models, often with a stated aim of achieving human or superhuman level artificial intelligence (see e.g. Altman, 2023), a number of smaller actors are developing a range of more specialised foundation models, some of which do not rely on large language approaches, for more specific purposes. These include models which specialise in mathematical word problem-solving (Mitra et al., 2024), models trained for the financial domain (Wu et al., 2023), and models specialised in single-cell biology (Cui et al., 2023).

This diversity could be taken to justify arguments that defining a single foundation model market or industry is just as problematic for analysis as the over-stretched concept of an AI industry is. Two key factors justify retaining the foundation model focus. First, the trajectory of general-purpose versus specialized models is uncertain, making it useful to have a flexible analytical scope. If general-purpose models achieve high performance at competitive prices, they may reduce the need for specialized models, leading to a more unified and identifiable general purpose foundation model market. Equally, the possible coexistence of multi-dimensional general-purpose models and specialized models which compete for some market segments highlights the need for a flexible analytical approach. Second, foundation models, whether general-purpose or specialized, share common inputs (compute, labour, and data), production methods, and downstream functionality as building blocks for applications. This commonality justifies a broader focus on the foundation model industry as a whole. Adopting this inclusive approach aligns with arguments that competition policy should evolve beyond traditional market-focused paradigms to better address the complexities of modern digital markets (Crane, 2022).

II. Literature Review

Despite widespread concerns about the industrial organization of the AI industry, which often implicitly refer to foundation models, deeper analytical examinations, particularly from academia, remain limited. Existing studies on competition in the industry have primarily approached it from a policy perspective, limiting the depth of their industrial organisation analysis to the identification of economies of scale and scope and speculation about possible network effects. They have typically examined the entire value chain without first isolating and analysing the dynamics characterising and emerging from foundation model production. This is in contrast to the cloud computing industry which has received extensive attention both independently (Benzina, 2019; Newsome, 2020; Ofcom, 2023) and as a crucial input for AI industries (e.g. Belfield & Anonymous, 2023). The model level, in many ways the industry's nucleus, has not been subject to focused analysis in this way, a gap this paper seeks to fill.

The most notable contributions explicitly focused on foundation models to-date come from the United Kingdom's Competition and Markets Authority (CMA) which published two reports examining the current and future state of competition in the industry (CMA, 2023a, 2024a), and France's Autorité de la concurrence which published its own equivalent in June 2024 (Autorité de la concurrence, 2024). Though their economic analysis is limited by virtue of their nature as policy reports, both are valuable contributions for the rich industry insights they provide having interviewed a range of stakeholders to inform their findings. The CMA report implicitly engages in natural monopoly analysis, but only at a superficial level. It notes the presence of economies of scale and scope, and indirectly recognises the importance of specialisation, suggesting that consolidation is most likely in a future in which models are extremely large, and the only competitive models are high-performing general-purpose cutting-edge products with extreme sunk costs (CMA, 2024a).

Vipra and Korinek, conversely, explicitly address the question of whether foundation models are "natural monopolies or oligopolies" (Vipra & Korinek, 2023: p.3). They answer in the positive, identifying the presence of economies of scale and scope, network effects, and limited access to key resources as factors to support this. While their work includes a number of important findings supported by strong empirical observations, it does not engage deeply

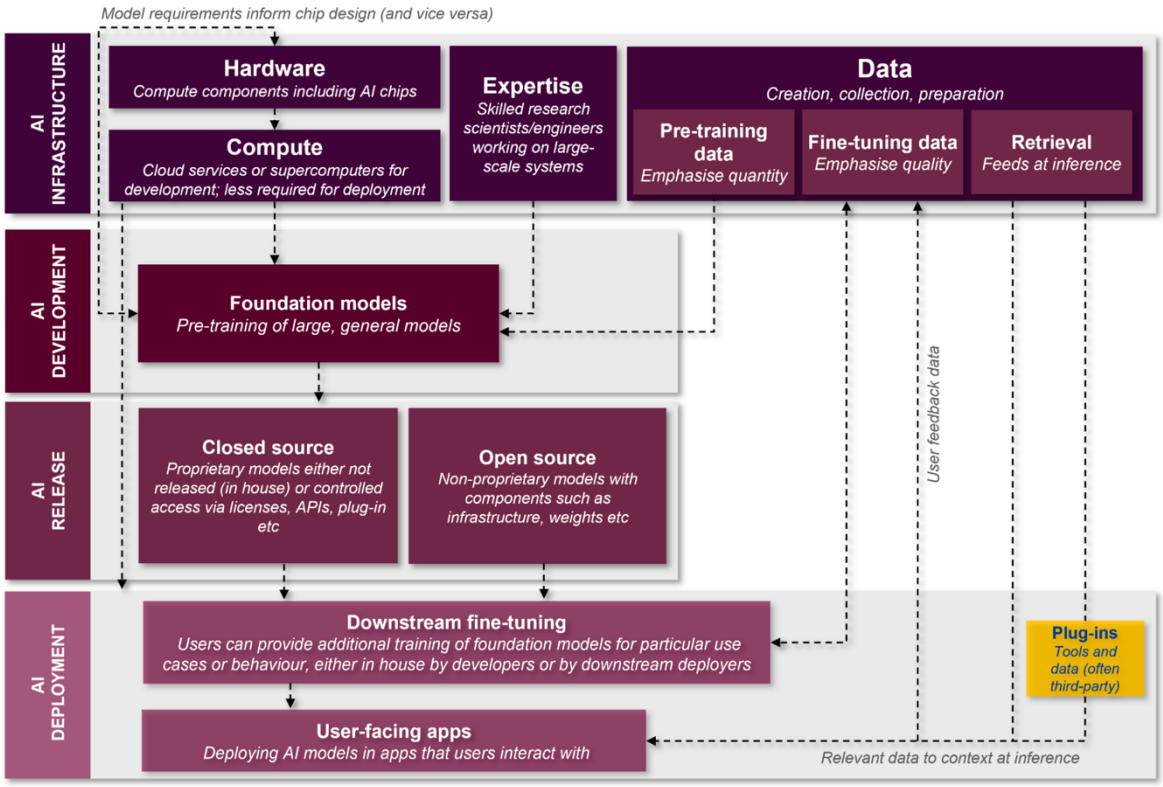
with industrial organisation theory. Vipra and Korinek do make one key contribution, however, which is to suggest that we might benefit from considering the 'cutting-edge' models and those "behind the frontier" separately, finding the market for the first to exhibit natural monopoly tendencies while suggesting the latter remain competitively dynamic (Vipra & Korinek, 2023: pp.37-38).

A more focused contribution which does draw on industrial organisation theory comes from Narechania (2021). He examines whether machine-learning-based systems might be natural monopolies, drawing usefully from the computer science, economic and legal literatures. He argues that where fixed costs of development and the computational costs of optimization are high, and where feedback learning network effects are strong, it is "likely that, for at least some [machine-learning-based] applications, a natural monopoly exists" (Narechania, 2021: p.1570). These findings can readily be applied to foundation models, which are machine learning models by definition (Bommasani & Liang, 2021). Narechania engages more explicitly and specifically with the economic theory of natural monopoly than other contributions, though his focus remains regulatory rather than economic. He demonstrates that the cost structure of producing the highest performing machine learning models fulfils all the requirements of natural monopoly. He also, somewhat speculatively, argues that techniques such as continuous machine learning (CML), where models 'learn' from new data streams without being retrained, may enable models to benefit from network effects, generating demand-side natural monopoly dynamics (Narechania, 2021).

However, the wide scope of Narechania's work makes it challenging to translate from abstract theory into practice and thereby inform policy. The category of machine learning models is incredibly broad, encompassing everything from cutting-edge foundation models to specific, smaller-scale models that can be trained on a personal computer (Kelleher, 2019). This forces him towards the tentative conclusion that "some machine-learning-based applications may be natural monopolies", pre-empting a closer analysis of the consequences of these theoretical findings on markets and firm behaviour. Instead, he progresses immediately to the question of regulation, arguing that traditional natural monopoly regulation may be an answer to market failures flowing from the establishment of these theoretical natural monopolies. This fails to consider the fact that natural monopoly dynamics do not necessarily translate to natural monopolies, especially in industries with such rapid innovation, scope for differentiation, and complicating external factors.

A key concern identified in multiple analyses of competition in the foundation model industry is in the way its unique vertical configuration, often referred to as its 'stack' (see e.g. Tsaih et al., 2023), creates potential for concentrations of market power to be leveraged vertically. Layered, or 'stack' thinking represents a "useful heuristic in critical examinations of mobile computing systems" (White, 2016: p.132). In the foundation model case, it helps provide a framework for studying the complex vertical relationships between different actors and layers of the industry's value chain. These can be uni- and bi-directional, dependent or co-dependent, creating opportunities for power to be leveraged across layers (see Fig.1). Models of the foundation model 'stack' are commonly used to underline how concentration at one level in one part of the industry, can have harmful up- or downstream effects on competition.

Figure 1. The foundation model industry framed as a stack



CMA (2024)

The CMA report identifies three risks to fair competition, all of which focus on these vertical concerns. They are critical input foreclosure, refusal to deal, and attempts to achieve these aims through partnerships while circumventing merger regulation (CMA, 2024). Vipra & Korinek (2023) identify similar concerns, and although they extend their discussion to

consider pricing strategies and predatory pricing, the dynamics and consequences of wars of attrition are not discussed. Belfield & Anonymous (2023) focus specifically on potential issues relating to merger control, abuse of dominance, state aid, and anticompetitive agreements in compute, focusing on their possible downstream effects on competition in AI, including foundation models. Similarly, although Hoppner & Streatfeild (2023) recognise the potential for concentration "at the AI modelling level" specifically, their analysis considers this alongside concentration at the compute, data creation, and application development levels. The harms they identify are also focused on vertical integration, upstream and downstream leveraging of dominance, and unequal access to inputs. As such, no study has yet focused specifically and comprehensively on the competitive dynamics intrinsic to foundation model production.

While it is crucial to consider vertical concerns, conducting such analysis without first considering concentration dynamics at each level in isolation is premature. Without a comprehensive independent understanding of the dynamics emanating from, and characterising, each level of the stack, the development of effective targeted policy responses to address market failures is hindered. This is especially true for the foundation model development level which forms the nucleus of the industry. This paper aims to fill that gap in the literature by contributing a focused industrial organisation analysis of foundation model production, examining to what extent its attributes give rise to natural monopoly dynamics. Beyond identifying their presence, which others have also done, albeit less rigorously, it also considers their dynamic consequences. The findings of this study are then integrated into the regulatory debate, where their implications for vertical concerns are examined.

While the natural monopoly question in foundation models development has not been extensively analysed, the same cannot be said for its equal-most-significant input - cloud computing services. Niyato et al. (2009) have modelled optimal strategies for competing firms based on the natural market structure of cloud computing. More recently, Benzina (2019) and Newsome (2020) contribute articles firmly influenced by natural monopoly theory, arguing cloud computing should be considered and regulated as an essential facility. Similarly, Ofcom acknowledges natural "limits to the overall level of competition" in the cloud computing industry (Ofcom, 2023: p. 9). While this emphasises the importance of considering vertical influences on competition in the foundation model industry (Belfield & Anonymous, 2023), it does not negate the need for comparably focused and theoretically grounded analyses of foundation model production.

III. A Very Short Introduction to Natural Monopoly Theory

The intellectual history of natural monopoly theory goes back to the 19th century (Mosca, 2008; Sharkey, 1982). J.S. Mill already explored idea that different markets and industries might naturally tend towards different degrees of concentration, sometimes resulting in monopolies 'created by circumstances, and not by law' in 1848 (Mill, 1848: p.483). The neoclassical economists who sought to understand these 'undertakings which are monopolies by virtue of their inherent properties' (Ely, 1894: p.294) for a long time focused entirely on their technologies of supply. They examined the causes and limits of economies of scale and the theoretical question of whether scale economies and perfect competition are incompatible (see e.g. Edgeworth, 1911; Knight, 1921; Marshall, 1898; Pareto et al., 2020; Sraffa, ([1925]1998); Stigler, 1951, 1957).[5]

Though broadly accurate, the neoclassical model is uncomfortably static. Baumol, Bailey, Panzar and Willig refined it, demonstrating that economies of scale across the relevant level of demand are sufficient, but not necessary, for natural monopoly (Baumol, 1977; Baumol et al., 1977). A firm producing a single homogeneous product is a natural monopoly when it is less costly to produce any level of output of this product with a single firm than with two or more firms over the full range of market demand (Baumol et al., 1982). When this is the case, we say the firm's cost function is subadditive. This can be expressed mathematically and graphically. A single firm's cost function is subadditive when it can produce all units at a given level at a lower cost than the sum of any multiple k companies with the same cost functions producing at the same total level. Taking $Q = q_1, \ldots, q_k$ to represent a vector of outputs in a market and C(Q) to represent the cost of inputs required if q is to be produced, subadditivity exists for a single-product firm when:

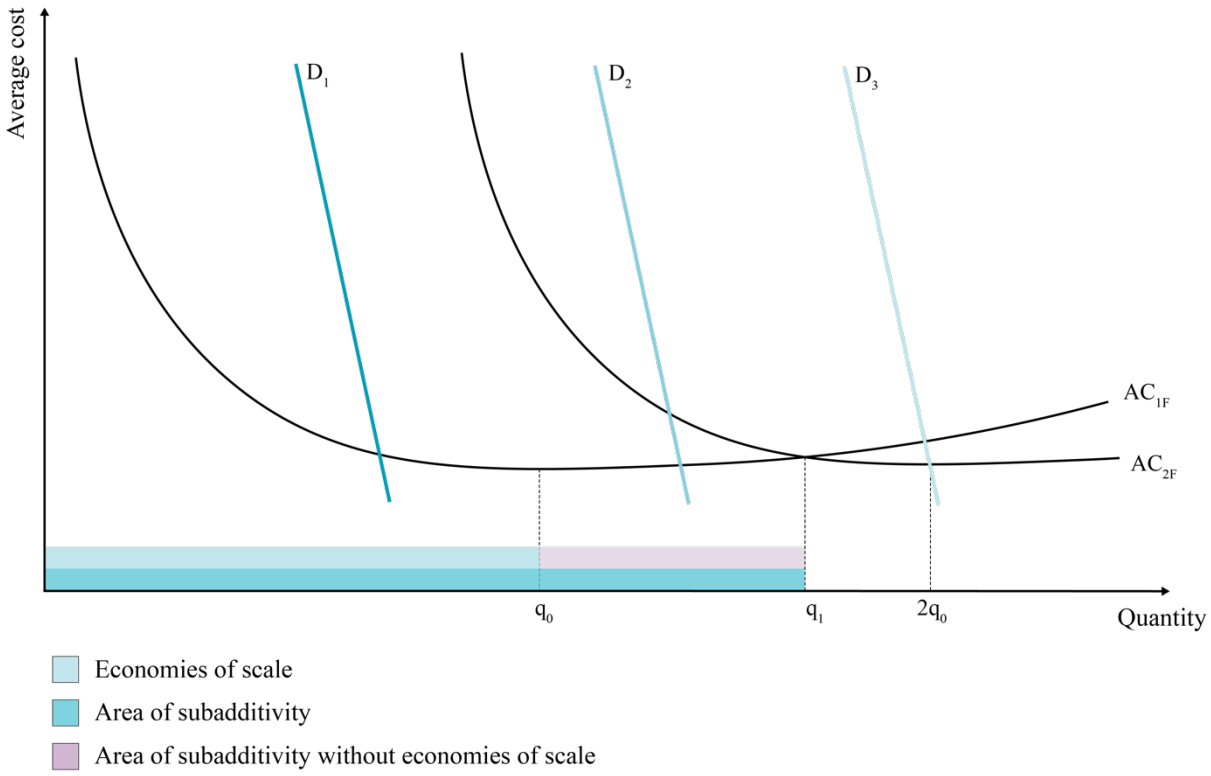
$$C(Q) < C(q_1) + \cdots + C(q_k)$$

Crucially, it can be less costly for output to be produced by a single firm even beyond the point where that firm experiences economies of scale, where average cost per unit decreases as production increases due to fixed costs being spread over more units. Fixed costs are expenses that do not change with the level of output produced, such as rent, salaries, or in our

[5] For an excellent deeper overview of the history of the theory of natural monopoly, with an Italian flavour, see Mosca, (2008)

case the cost of training a model. Marginal cost, the additional cost incurred to produce one more unit of output, may increase as production scales. Average cost (AC) is calculated by dividing the total cost by the number of units produced, representing the cost per unit. The concept of subadditivity is illustrated in Figure 2 where, despite the firm experiencing diseconomies of scale from q_0 as marginal cost increases, it remains more efficient for one firm (AC_{1F}) than two (AC_{2F}) to produce in the market up to point q_1. Subadditivity thus tells us that increasing marginal costs are not incompatible with natural monopoly. Baumol also illustrated and modelled an extension of this theory to the multiproduct case to demonstrate how natural monopolies can be achieved through economies of scope, wherein it is most efficient and cost effective to produce a combination of multiple products (Baumol, 1977).

Figure 2: Cost curves of two firms demonstrating the concept of subadditivity



Everything discussed up to this point can be referred to as the technological, or cost-based, definition of natural monopoly (Joskow, 2007). While this determines whether a market is a natural monopoly market, it does not tell us *what kind* of monopoly market it is. But that detail is decisively important, especially for regulatory purposes, because the implications of

a natural monopoly being sustainable, contestable, temporary, or networked differ considerably. Game theoretical approaches examining the strategic behaviour of firms in the presence of natural monopoly or oligopoly cost structures allow us to model these markets dynamically. They help identify not only why natural monopolies exist but also why and how they are sustained, and how they can give rise to problematic behaviour and situations.

High fixed costs and economies of scale alone do not deter entry in a way that breeds sustainable natural monopolies with monopoly power. For that to happen a significant proportion of fixed costs must be sunk. That is, once they are incurred, they cannot be recovered, regardless of future actions (Sutton, 1991). Indeed, the Baumol group's most significant contribution to natural monopoly theory was not subadditivity, but their recognition of the fundamental importance of sunk costs (Baumol et al., 1982).

Sunk cost considerations provide the linkage between subadditivity and game theoretical behavioral definitions and analyses of natural monopoly. As the primary determining feature of incumbents, sunk costs create an asymmetry between firms that are 'in' the market and potential entrants. They are 'what make the distinction between incumbents and potential entrants meaningful' (Joskow, 2007: p.1245). Once incurred, sunk costs no longer factor into incumbent firms' strategic pricing decisions because their entry is committed: their decisions only consider marginal costs. They however represent a significant barrier to entry as potential entrants will not enter a market if they do not think its prices will allow them to recover these costs. The more significant the sunk costs, the more sustainable the monopoly. This creates opportunities for incumbents or first movers to behave strategically in specific ways that give their monopolies stability and sustainability, minimising the contestability of the markets they monopolise. Understanding and pre-empting such behaviour where it creates artificial barriers to entry is essential to managing the risks that come with natural monopoly. Therefore, the key requirements for sustainable natural monopoly, or oligopoly, to exist, therefore, are substantial increasing returns combined with long-lived sunk costs that represent an significant fraction of total costs (Joskow, 2007; Sutton, 1991).

Baumol, Panzar, and Willig's theory of contestability demonstrates how competitive outcomes are theoretically possible in some natural monopoly markets. The threat of competition for the market, they argue, can have almost the same influence on prices and incentives as competition in the market because the monopolist has to price at average cost. Three major conditions must exist for a monopolised market to be absolutely contestable: no

barriers to entry or exit; the possibility of hit-and-run entry where competitors can enter the market, compete, and exit without losing their investment if they cannot sustain profits; and no, or very low, sunk costs (Baumol et al., 1982). Contestability theory has proved highly influential, notably underpinning arguments in favour of deregulation in the 1980s. It also influenced some of the thinking behind the DMA whose alternative title is "Regulation on contestable and fair markets in the digital sector". It seeks lower barriers to entry in digital markets by imposing obligations on gatekeepers controlling key market infrastructure and by mandating interoperability in some areas. Contestability theory is however not free from criticism, notably because the conditions of no entry and exit barriers and no sunk costs are rarely met in real-world markets (Brock, 1983; Weitzman, 1983). While it should therefore only justify non-intervention where its conditions are for the most part met, it underlines the importance of making natural monopoly markets as contestable as possible by minimising barriers to entry.

In rapidly evolving industries, naturally monopolistic cost structures do not necessarily result in natural monopolies. Though static models may predict them, dynamic ones help to explain why they do not necessarily materialise or only do so temporarily. Instead, the demonstrate how these cost structures can generate natural monopoly dynamics, which give rise to different competitive concerns and market failures. This is especially the case in technology markets chracterised by rapid innovation. Game theoretical approaches examine the strategic behaviour of firms operating under these conditions. These include the study of dynamic pricing strategies used to manage demand and deter entry while responding to regulatory constraints (Milgrom & Roberts, 1982; Tirole, 1988); the study of strategic preemptive sunk investments (Fudenberg & Tirole, 1984; Milgrom & Roberts, 1982); and 'war of attrition' models (Bulow & Klemperer, 1999; Tirole, 1988).

Finally, network effects, which occur when the value of a product or service increases as more people use it (Katz & Shapiro, 1985), can breed a unique form of demand-side natural monopoly. Game theoretical approaches naturally lend themselves to the modelling of market competition in the presence of network effects (Katz & Shapiro, 1994; Rochet & Tirole, 2003). Demand-side network effects can introduce a cost-independent source of natural monopoly by creating increasing returns to the scale of demand. They create behavioural barriers to entry for potential entrants. In our connected present, it is always necessary to examine whether network effects are present when examining one industry's industrial

organisation. This is especially the case in many digital markets such as search, social media, and online marketplaces where the presence of extremely powerful network effects is the main force driving them to demonstrate natural monopoly dynamics (Ducci, 2022). An important question for our purposes is whether network effects will play a significant role in the foundation model industry, and if so how that might influence monopoly dynamics.

IV. An industrial organisation analysis of foundation model production

A. Network effects

Compared with many digital markets where they are the primary force driving force natural monopoly dynamics, network effects are unlikely to be as significant in the foundation model industry. First, there are no direct network effects like the ones that drive concentration in markets such as telecommunications and social media networks (Katz & Shapiro, 1985). Neither are there strong indirect or cross-side network effects like the ones that drive concentration in two-sided markets like search engines and digital marketplaces (Rochet & Tirole, 2003). The only possible network effects are data network effects, which occur when the value or utility of a product or service increases as more data is collected and used within the system (Gans, 2024; Hagiu & Wright, 2023). But these are only significant sources of competitive advantage when the data collected from users meaningfully improves that service, for instance in the case of search engine sorting algorithms where value is derived from better predicting user preferences. The relationship between user data and foundation model performance is much less direct than it is in the context of most existing digital monopolies because the tasks (Ducci, 2022). While data is a crucial source of competitive advantage in foundation model production, data generated by users' interactions with models is only likely to contribute to marginal gains in product improvement because it is not the type of data that their performance fundamentally depends on. If networks effect are present, their influence is 'relatively minor in comparison to the economies of scale on the supply side' (Vipra & Korinek, 2023: p. 11), which we turn to now.

B. Assessing the market structure of foundation model production

Market structure analysis begins with cost analysis. This involves examining the fixed and marginal costs involved in production, and the extent to which fixed costs are sunk. Foundation model production requires three major inputs: training compute, R&D labour costs, and data (CMA, 2024a). The nature of each and their relationship with production are considered in turn. This helps to determine whether cost curves are subadditive, and, if they are, the extent to which the resulting natural monopoly is sustainable. Product homogeneity is assumed for this stage of the analysis, but differentiation is considered and introduced into the model in section E.

Costs in foundation model production

Fixed costs

The cost to train the highest performing foundation models is extraordinarily large and continues to increase having grown at an exponential rate of 2.4× per year between 2016 and 2024. The most expensive of these cost over USD100m to train by the end of that period (Cottier et al., 2024). Two factors are likely to affect costs: changes in the cost of inputs, or changes in production methods resulting in more efficient use of inputs. Whether these fall, plateau, or rise will have an impact on the industry's market outcomes and competitivity. However, as the Baumol group illustrated, if costs decrease but the cost function remains subadditive, the market dynamics will remain naturally monopolistic.

Compute

Compute is the single largest cost in foundation model production. In 2024, the highest performing models' compute costs alone ran in the tens, and sometimes hundreds, of millions of US dollars (HAI, 2024). Without fundamental technical changes to production methods, exponential increases in computing power are necessary for linear improvements in model performance, with current estimates suggesting a quadratic relationship between training compute costs and the number of parameters (Faraboschi et al., 2024; Thompson et al., 2021). The massive increase in demand for accelerated computing, which uses advanced processors (mostly graphics processing unites (GPUs)) in conjunction with traditional central processing units (CPUs), caused bottlenecks, waiting times, and shortage-driven increases in

accelerated cloud computing prices and data centre construction costs between 2020 and 2024 (Hille & Liu, 2023). This partially explains the extremely high costs incurred to train models during that period.

Response to this demand shock may reduce the cost of compute in the medium-to-long term. Indeed, computational price-performance rates in GPUs have grown rapidly, doubling every 2-3 years between 2010 and 2024 (Hobbhahn et al., 2023). Demand has simply outpaced it. If the computational demands of training models stabilise or decrease through efficiency improvements or changes in approach to model development, fixed costs may therefore decrease substantially. Though increasing scale has dominated technical development in leading foundation models to-date (see e.g. Halevy et al., 2009; Hoffmann et al., 2022; Kaplan et al., 2020), a significant driver behind that trend appears to be the loftier goal of producing general artificial intelligence (AGI) as opposed to a foundation model with a certain set of functions that can serve a market demand. As markets for foundation model services develop, it is conceivable that specialised foundation models whose computational demands are stable, or at least grow more slowly than the costs of compute decrease, become widespread (see e.g. Chen et al., 2017; Lee et al., 2020; Poplin et al., 2018). The computational cost of training these more specialised models are magnitudes lower than those of the leading general-purpose ones, but can still run in the tens of millions (see e.g. Wu et al., 2023). This becomes relevant later when we consider different possible industry futures.

Labour

While compute costs are often taken as shorthand for total fixed costs, R&D staff costs have been estimated to account for 29-49% of total frontier foundation model development costs (Cottier et al., 2024). Like compute costs, R&D labour costs have been significantly influenced by shortages which may be remedied, or at least improved, in the medium-to-long term. They may therefore stabilise at lower levels as more talent is trained and attracted to the industry. Their extreme values in the early 2020s stems from the fact that by some estimates only a couple of hundred people possess the skills required to design and train the most advanced foundation models (Bindley, 2024). Firms not seeking to produce AGI-chasing models may also benefit from lower labour costs, it being less imperative for them to pay premiums for the highest tier of research talent. This is not a given, however, as talent likely remains a key source of competitive advantage.

Data

The future cost of data is unpredictable, and may well increase, particularly as new sources of public data become increasingly rare and competition for access to high-quality proprietary datasets intensifies as it becomes an important source of competitive advantage (Brown et al., 2020: p. 8). Producers are increasingly looking to access high quality data to train models, investing extensively in dataset development and signing a growing number of licensing deals with media and social media companies often worth several million dollars (Gilbert, 2024; Mauran, 2024). Among other factors, the cost of data may hinge on the outcome of ongoing lawsuits against foundation model producers, notably the one brought by the New York Times against OpenAI and Microsoft in late 2023 (*The New York Times Company v. Microsoft Corporation, OpenAI, Inc.*, 2023). These may force producers to pay significant fines or compensation for data they have already used and set a precedent which makes data substantially more costly. With limited sources of high quality training data available, these are also liable to become a coveted commodity with correspondingly high prices as producers seek exclusive deals for them (Gunasekar et al., 2023; Li et al., 2023; Villalobos et al., 2024).

Sunk costs

"[T]he extent to which capital is sunk determines the monopoly power and profit enjoyed by established firms" (Tirole, 1988, p. 388). It is therefore important to carefully consider what proportion of the fixed costs involved in foundation model production are sunk. R&D and model training costs (notably computational costs) are unrecoverable. While investments in GPU-heavy infrastructure made by firms that do not outsource their computing needs might be partially recoverable through repurposing, the money spent running and powering these computing infrastructures for training models is sunk. For producers who outsource their computational training needs to third parties, all of those costs are sunk. Investments in dataset curation, conversely, may be partially recoverable given the high demand for high quality datasets. Overall, however, the fixed costs of foundation model production clearly involve an exceptionally high proportion of sunk costs.

Marginal costs

Deploying foundation models incurs the non-negligible marginal cost of inference, which is the computational cost of using a model to make predictions or generate outputs after it has been trained (Patterson et al., 2022). However, inference is not a marginal cost of foundation

model *production* because it affects the consumer only after they have purchased access to the model. Some producers may adopt a business model whereby they market models by bundling them with inference computing, but that represents a different product.

The marginal cost of foundation model production is therefore its cost of distribution, which is low. This predominantly includes costs associated with hosting and storing the model to make it accessible to users (often via deals with cloud service providers), and the provision of customer support for users. Helpfully, these costs are similar regardless of whether a model is distributed using an open- or closed-source approach. Thus, compared with the fixed cost of model development outlined above, the costs of model distribution are thus minor. Moreover, even if they do at some point begin to increase, the theory of subadditivity tells us that this by no means prevents the market from being naturally monopolistic. Being several orders of magnitude lower than fixed costs, they would have to increase dramatically to materially affect average costs.

Foundation models are theoretical natural monopolies

Consider a single-product firm which produces a foundation model which outperforms all competitors when it is first marketed. As we have seen, foundation model production involves high fixed costs associated with initial R&D and training. While these may decrease through efficiency gains and technical advances in production, R&D investment and computational scale are intractably linked with foundation model performance and quality (Halevy et al., 2009). These necessary significant investments in data collection and preprocessing, model architecture design, and model training are indivisible: they cannot be scaled down without losing their utility (Edgeworth, 1911). As established above, the fixed costs of such a model are almost entirely sunk.

Meanwhile, the digital nature of foundation models means they have comparatively negligible marginal costs because distribution, the main variable cost, only requires the licencing and duplication of the trained model. Because digital distribution is virtually costless regardless of distance, it can also be global. Taken in isolation (i.e. discounting foundation models bundled with inference computing), such a foundation model's cost function demonstrates virtually unbounded economies of scale. At a fundamental level, foundation model production is thus subadditive across all relevant quantities.

Importantly, this finding applies regardless of the business model adopted by a producer. Even if foundation models are marketed for free through open-source distribution and licencing, with the seller recouping the cost of production through bundling or by offering a complementary product, the most efficient means of satisfying market demand is still through monopoly production. Strictly-speaking, an open-source approach can be understood as an extreme form of predatory pricing (Areeda & Turner, 1975) which removes any possible incentive to produce a close substitute product, creating further natural monopoly dynamics.

The key requirements for natural monopoly, significant increasing returns combined with long-lived sunk costs that represent a significant fraction of total costs, are thus inherent to foundation model production. If technology is constant and the product is homogeneous, therefore, foundation models are natural monopolies. No producer of a substitute product could, or should, attempt to compete if someone has already committed the up-front investment to train a model. In a static foundation model market, entry is socially undesirable and productively inefficient. It results in higher prices as the burden of duplicated costs are spread over the same number of buyers, and represents a significant opportunity cost (Edgeworth, 1911). As the next section which focuses on dynamics demonstrates, however, this static analysis – which is where previous analyses of competition in foundation model production end – provides an incomplete picture. It demonstrates that natural monopoly dynamics are present and influential, but it does not determine how, whether or why natural monopolies emerge, nor what they entail.

C. The dynamic consequences of the natural monopoly characteristics of foundation model production

Competition for markets through wars of attrition

Though foundation models have static natural monopoly properties, in a dynamic context intense competition can emerge as firms fight over the monopoly position. In new industries or markets like the foundation model industry which are recognised to exhibit natural monopoly tendencies, and therefore monopoly rents, but which are young enough that no monopolist is yet installed, war of attrition scenarios commonly develop (Posner, 1975). As opposed to competition in markets, where firms compete for market shares, this involves competition *for* markets, where firms compete to establish dominance and secure exclusive control over the entire market (Geroski, 2003). These arise particularly often in battles to control new technologies as firms compete intensely to benefit from long-term monopoly rents that often span global markets (Bulow & Klemperer, 1999).

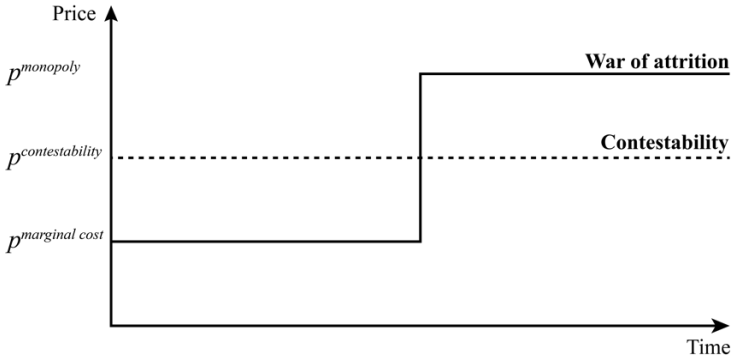
In a Betrand competition market, where firms producing a homogeneous product compete on price, firms are incentivised to endure losses for a certain amount of time because of the substantial returns promised if they achieve the monopoly position. In game theory, this situation is modelled as a continuous-time game where firms make strategic decisions about whether to exit the market based on estimations of their own and their competitors' financial endurance. Firms anticipate short-term losses to avoid losing their sunk costs and continue competing, leading them to undercut each other to marginal cost. This results in zero economic profit until all firms bar one drop out, whereby the winner raises prices to monopoly levels (see Figure 2). An initial equilibrium is achieved while the war is ongoing, with prices driven down to the marginal cost of production. With multiple competitors sinking costs to compete in a market that is productively efficient when only one firm supplies the market, redundant facilities are built and maintained (Tirole, 1988).

Game duration (i.e. how long multiple firms compete at marginal cost pricing) depends on industry features, increasing as the value of the prize increases but decreasing as the costs necessary to stay in the game increase (Tirole, 1988). Indeed, if commitment is sufficiently high and operational costs are sufficiently low, indefinitely long games are possible. This is especially true in the presence of wealthy sponsors able to bear and subsidise the cost burden.

Examining the efficiency consequences of this, it is important to note that there is no single measure of efficiency in welfare economics. It is necessarily a 'value-laden' and normative term (Blaug, 2001, p. 47). Three key measures are commonly used. A market is Pareto efficient when resources are allocated in such a way that no individual can be made better off without making someone else worse off. Allocative efficiency is achieved when resources are distributed such that they maximize total societal welfare. Finally, productive efficiency occurs when the market produces at the lowest possible cost.

In wars of attrition, none of these is achieved, primarily due to the wasteful duplication of costs. Though marginal cost pricing increases consumer surplus while the war is underway, this is inefficient by all three measures as firms incur unnecessary duplicated costs and operate at a loss. After the war is won, the market inherits the classic inefficiencies of a monopoly market. Though productive efficiency can be achieved this way, allocative and Pareto are not as monopoly prices and quantities cause welfare loss. Figure 2 compares a war of attrition situation with the price in a contestable natural monopoly market, which Tirole (1988) convincingly argues is the social planner's preferred outcome in natural monopoly situations as it is the lowest possible price (P = ATC) at which the firm is willing to operate.

Fig.2 – Prices in a war of attrition



Wars of attrition: constant and homogeneous products

Consider first a foundation model market where the product is constant and homogeneous (i.e. assuming innovation has plateaued, making all models close substitutes). As we have established, the market structure in this situation is natural monopoly. A war of attrition for that foundation model monopoly is highly likely given the expected value of the prize and the

industry's immaturity. As of 2024, anecdotal evidence supports this prediction, with multiple players including OpenAI, Google, Meta, Mistral, and Anthropic competing intensely, providing general purpose foundation models at below-cost prices or even for free, and failing to recover costs (Efrati & Holmes, 2024).

Game duration is significant in this context. Exceptionally high prize expectations combined with very low marginal costs mean that the benefit of staying in the game diminishes very slowly. Prolonged games are therefore expected, with the significant inefficiencies that accompany the war phase accumulating over a longer period. That said, inefficiencies in duplicated foundation model production primarily occur at the point of market entry, when multiple firms commit to the significant unrecoverable sunk costs of developing their models. While some facilities can be repurposed, the nature of foundation models means that the majority of capital spent—including all R&D and training compute costs—is irretrievable.

Because inefficiencies are front-loaded as sunk costs, extending the competition based solely on cost and price is less problematic than one might intuitively expect. Regardless of the game's length, the inefficient investment was committed at entry, so prolonging it does not create additional inefficiencies. However, if firms compete through other means, such as non-price competition (e.g. spending on marketing, legal fees, and lobbying), the worst-case scenario could occur. In such cases, all potential monopoly rents could be eroded through wasteful expenditures, making a longer war highly destructive (Joskow, 2007).

Wars of attrition: the dynamic model

While the analysis presented in the previous sections provides important insights into the fundamental monopolistic dynamics that characterise foundation model markets, it rests on two major assumptions: product homogeneity and no innovation. Innovation is introduced to our model to better reflect reality, examining whether and how this influences competitive dynamics. With no incumbent, high sunk costs, and a prize of monopoly profits, a war of attrition is again likely. In this game, however, the war is fought across two dimensions: price and innovation.

This can be modelled by modifying the standard war of attrition model to incorporate innovation. Firms in this scenario no longer only decide whether or not to continue

competing and incurring marginal costs as in the standard war of attrition. Instead, alongside deciding whether or not to continue incurring marginal costs, the game is played in rounds, with players also having to decide at each round whether they incur another round of sunk costs to innovate and train a new model that achieves the next level of performance. Provided firms innovate approximately simultaneously and to the same level, the potential competitive gains from innovation cancel each other out. Firms will continue to innovate purely because not doing so represents a loss of the chance to win monopoly profits. However, marginal cost pricing must also continue because, products remaining substitutes, firms also undercut each other as they compete on price. The outcome is therefore similar to the standard war of attrition model, the main difference being that the choice to incur another round of sunk costs each time represents a more significant decision for the players given the costs involved.

The cost of continuing to play this game is much more significant than it is in the static product case where only low marginal costs are incurred. High prize expectations continue to increase expected game duration, with the value of the prize possibly even increasing if product improvement generates higher demand. But the cost-benefit analysis of staying in the game is changed by the need to incur repeated high sunk costs to continue competing. The higher the cost of staying in the game, the faster a player's expected benefit from continuing to fight diminishes, and the quicker they ought to leave the market (assuming they act rationally and do not fall victim of the sunk cost fallacy).

From an efficiency perspective, the introduction of value-generating innovation alters the picture significantly. It is largely similar to the original static model in the way duplicated costs engender productive, allocative, and Pareto inefficiencies during the war, with only productive efficiency likely achieved at its resolution. However, the value generated by the innovative activity driven by the war for potential monopoly profits adds a new element to consider. This calls for a dynamic understanding of efficiency which considers the optimal allocation of resources over time, with a Schumpeterian focus on the importance of innovation and technological progress to enhance long-term welfare and growth (Blaug, 2001).

While the war is fought and price remains equal to marginal cost, the value added by innovation-driven product quality improvements is handed over in its entirety to the consumer, continually increasing consumer surplus. Though a monopoly or oligopoly may still emerge, with prices rising accordingly, the welfare benefits generated by the innovative

competition for the market can be consequential. As such, the presence of duplicated costs might still be considered dynamically efficient so long as that duplication results in a welfare-generating innovation payoff which exceeds the value of these costs. This is in line with the Schumpeterian defence of monopoly (Schumpeter, 1944).

From a welfare perspective, ideal game duration in this dynamic situation is more complex. Though an endless war maximises consumer benefit, societal dynamic efficiency is maximised if the war endures while innovation payoffs are large but ends if or when they plateau. If we assume that innovation follows an S-curve whereby it starts slowly, accelerates rapidly, and eventually plateaus as improvements become marginal (Kucharavy & De Guio, 2011), a dynamically efficient outcome is achieved if facilities are only duplicated while the cost of doing so is less than the value generated by the innovation that accompanies this duplication. Continuing to fight via innovation when the innovation returns on investment become negative is only marginally better than situations where firms compete away all potential monopoly rents through non-price competition (Joskow, 2007).

This game demonstrates how a war of attrition fought through cost and innovation multiplies duplicate costs while simultaneously spurring innovation, a trade-off policymakers must consider, especially in (at least initially) rapidly innovating industries like the foundation model one. It also illustrates how the long-term cost of innovation in foundation models (the sum of the repeated sunk costs) contributes to the final monopoly price consumers will have to pay.

D. Contestability

Contestable markets theory says that even in natural monopoly markets, the external threat that someone might enter and compete for the market can have almost the same influence on prices and incentives as competition in the market does (Baumol et al., 1982). This is unlikely to be the case in foundation model markets if and where monopolists or oligopolists emerge. High sunk costs and other barriers of entry, quick price responses and the ability to engage in above-cost limit pricing all imply a firm limit on the extent to which foundation model markets can be sufficiently contestable to prevent monopoly market failures.

High sunk costs mean there is no incentive for new firms to enter and incur the substantial expenditures required to train a foundation model and compete because they cannot recoup them: they know entry will not be profitable. Hit-and-run entry is also impossible because price responses can be rapid given the nature of digital production. New firms therefore cannot undercut the monopolist to steal their business and make significant profit for any relevant amount of time because incumbents can adjust immediately respond by cutting their own price. This further lowers chances of contestability. Without a realistic threat of entry, the standard monopoly pricing market failure and its associated welfare costs are therefore to be expected in foundation model markets.

Despite this, maximizing contestability is still important as it can still generate some competitive pressure, influencing the magnitude of these market failures. Absolute decreases in sunk costs, though they do not make the market absolutely contestable, push market dynamics in that direction. Lowering barriers to entry and exit generates more competitive pressure on incumbents. Even if they can sustain their monopoly and charge higher-than-optimal prices, more contestability generates more price discipline and incentives to be efficient and innovative. Even though the incumbent can charge a price higher than under contestability, greater potential competitive threats force them to engage in strategic pricing strategies –for instance pre-emptive or limit pricing– that result in their charging prices lower than the monopoly price (Bain, 1949; Milgrom & Roberts, 1982). The lower prices act as a deterrent, signalling to potential entrants that the incumbent is prepared to drive prices down if entry is attempted. At the same time, though, they also result in consumers being charged a price closer to welfare-maximising levels. The monopolist is not fully disciplined but may be more disciplined.

Higher contestability amplifies competitive effects because incumbents must remain vigilant against new entrants offering cheaper and more innovative products. This also incentivizes incumbents to keep innovating and prevents them from raising prices too high. Regarding innovation, potential free riders pose a challenge. Firms entering the market later incur fewer rounds of sunk costs compared to earlier entrants. For instance, a firm entering the market now doesn't need to recoup the costs from previous outdated models, allowing it to compete at lower prices. While established firms benefit from learning advantages, R&D investments, and established consumer bases, later entrants can avoid the high costs incurred by early innovators, giving them a competitive pricing edge.

E. Differentiation: specialisation, homogeneity, and multi-dimensional competition

Until now, our models of foundation model competition have assumed product homogeneity. That is a possible but far from certain outcome. The industry is dynamic and rapidly changing, with some firms attempting to develop general purpose products able to be applied to a vast range of tasks while others are producing specialised models with narrow or domain-specific tasks in mind. The extent to which product differentiation will characterise and define this industry is an open and complicating question. This section addresses this by acknowledging that uncertainty, using it to frame findings and analysis by systematically considering the competition consequences of different specialisation outcomes.

A taxonomy of possible foundation model industry outcomes is considered, focusing on the main non-price source of competitive advantage: product differentiation (Porter, 1985). Differentiation in the industry can be understood across two key dimensions: homogeneity and specialization. Homogeneity refers to the extent to which foundation models are similar in terms of capabilities, design, and application. This is important because in a highly homogeneous market, foundation models are highly substitutable, leading to increased competitive pressure and a focus on standardization and incremental innovation. Specialization, conversely, pertains to how foundation models are tailored for specific tasks, domains, or user needs. It creates market niches, fostering innovation driven by specific segment needs. The following section examines how the natural monopoly dynamics implied by the technology's subadditive cost functions play out in different differentiation scenarios as set out in Table 1.

Table 1. Taxonomy of foundation model differentiation scenarios

	Homogeneous	Heterogeneous
General-purpose	• General purpose models able to perform highly across a wide variety of tasks. They are extremely versatile and can be adapted to tasks through fine-tuning. • Standardization occurs as best practices are adopted, resulting in largely similar models with minimal differentiation. • Price and one-dimensional performance competition.	• General purpose models able to perform highly across a wide variety of tasks. They are extremely versatile and can be adapted to tasks through fine-tuning. • Though general-purpose in nature, models vary in their approaches, architectures, and strengths, • Competition focuses on differentiating features, performance nuances, and user experience.
Specialised	• Models are specialized for specific tasks or industries but are similar within those specializations. • Competition is within niches, focusing on efficiency, cost, and minor performance differences.	• Models are tailored for specific applications and differ significantly in their designs and capabilities. • Competition focuses on innovation and diversity, with models highly optimized for particular use cases.

General-purpose/homogeneous

The first scenario envisions a future dominated by general-purpose models. This outcome, highly desired and prominently championed by leading foundation model producers and their major financial backers, sees general-purpose models characterizing the industry. From a policy perspective, this homogeneous general-purpose outcome warrants special consideration due to the substantial market size and value controlled by a monopolist, as well as the potential up- and downstream consequences of such concentrated power.

In the extreme case, technological innovations result in the development of general purpose models that match or exceed the capabilities of specialized models in their individual domains (see e.g. Nori et al., 2023). Scale and scope have been key drivers of foundation model performance to-date, enabling generalist models to achieve high levels of performance for specialised tasks (see e.g. Halevy et al., 2009; Hoffmann et al., 2022; Kaplan et al., 2020). These technological trends may therefore privilege a general-purpose scenario, especially if they yield economies of scope, making them more cost-effective than the specialised models they compete against in sub-markets (Baumol, 1977).

Unsponsored standardization processes might occur if best practices and optimal architectures emerge, leading producers to adopt similar technical production methods. This was observed with the extremely widespread adoption of transformer-based architectures in

foundation model development (Wolf et al., 2020). User demand for interoperability and ease of integration across platforms and services can push the industry towards uniformity, encouraging firms to converge on standardized solutions to maximize market acceptance and reduce development costs, resulting in a homogeneous product market (David & Greenstein, 1990).

Dynamics and outcomes in this case are expected to follow those set out in the preceding sections. With very high sunk costs, the market tends towards natural monopoly, with single-firm supply the most productively efficient outcome. Until an incumbent emerges, an innovation-heavy war of attrition occurs as firms initially engage in long-lasting innovation and price competition to maintain competitive advantage. Innovation welfare gains are initially entirely transferred to consumers. Lower costs may contestability by reducing barriers to entry and will extend consumer-benefitting price wars, but competitive dynamics are unlikely to fundamentally change given the primacy of economies of scale. Innovation may therefore slow, and prices will approach monopoly levels.

General-purpose/heterogeneous

The first alternative idealised scenario also has general purpose models dominating for reasons articulated above, but with scope remaining for product differentiation along the homogeneity dimension. Products remain heterogeneous with sufficient consumer desire for variety which cannot be served by a single model. Differentiation will occur if general-purpose models with different architectures, training methods, and datasets vary in task strengths and weaknesses (Center for Research on Foundation Models (CRFM), n.d.; Liang et al., 2023), price-performance ratios, or inference costs (Patterson et al., 2022; Villalobos & Atkinson, 2023). For instance, though both BERT and GPT-4 are general purpose large language models, their different architectures and training methods mean BERT models outperform GPT models on classification tasks while the opposite is true for text generation (Bosley et al., 2023).

If technological limits on model performance means consumer tastes are best satisfied by a range of general-purpose foundation models with different affordances and prices, firms have an incentive to cater to those. It follows that there will be entry to the extent that there is room for a variety of products, with the intensity of competition determined by consumers at the

margin between those products. So long as there is a sufficient degree of substitutability, the outcome is more competitive as producers compete on cost, reliability, and performance.

Although foundation model production demonstrates natural monopoly characteristics as demonstrated in Part IV.B, the emergence of natural monopolies is contingent on a certain degree of homogeneity (as with traditional public utility products including water and electricity). It is well-established that differentiation leads to intermediate prices and profits between those of monopoly and perfect competition by generating a new competitive dimension (Shaked & Sutton, 1982). Cost functions remaining subadditive, and barriers to entry high, the expected market structure is therefore more competitive, but only moderately so, especially if the scope for product differentiation is constrained. We therefore expect an oligopoly of firms to emerge with differentiated but overlapping offerings. Concentration may be amplified by economies of scope.

Specialised/homogeneous

A specialised-homogeneous scenario is likely to emerge if there are technological and commercial advantages to catering to specific demand niches. This will be the case if consumer demands are best served by distinct models specialising in different task categories, with general-purpose models neither more performant nor more cost-efficient compared to specialised alternatives. The market will thereby be segmented into niches where differentiated models excel in distinct, specialised domains such as translation, image generation, text summarisation, biomedical (Cui et al., 2023), and legal (Chalkidis et al., 2020) tasks. Unlike the general-purpose-heterogeneous scenario, the differentiation emerging due to specialisation is more defined as models are developed with specific rather than general use-cases in mind.

If homogenization occurs within these niches following standardization and best practice development, we expect natural monopolies to emerge in each sub-market. Wars of attrition are possible, but may be resolved quicker, especially in markets with smaller prizes – though lower costs might counteract this somewhat. The eventual outcome has strong natural monopoly dynamics if or when an incumbent gains control of each sub-market. Similar to the general-purpose scenario, the dynamics in a homogeneous product market will not be significantly affected by lower costs unless these are extreme. Outcomes and concerns are

similar to those in the general-purpose-homogeneous case, differing only in scale given the smaller markets involved. Because of lower absolute sunk costs in each sub-market, they are individually more likely to be contestable than a single general-purpose market.

Specialised/heterogeneous

Finally, we consider a scenario characterised by high degrees of specialisation and differentiation resulting from a combination of the technological and economic forces discussed in the general-purpose-heterogeneous and specialised-homogeneous scenarios. In this case, market demand is best served by specialised models, but the technical possibilities available to cater to those demands result in a heterogeneous range of products within each specialism. Within specialised domains, therefore, further differentiation emerges, for instance over price, performance, and the cost of inference.

In this outcome the market is fragmented and heterogenous as a diversity of foundation models emerge. They satisfy different needs as the technological limits and affordances of foundation model production preclude the vertical monopolies of the specialised-homogeneous scenario and the total monopoly of the general-purpose-homogeneous one. Of the four scenarios, this is the most competitive, with firms competing within various specialised niches, leading to a more fragmented industry with several sub-markets. The expected competitive outcome is a combination of the specialised/homogenous and heterogeneous/generalist cases, with a series of oligopolies likely to emerge in each specialised market.

Table 2. Competitive consequences of possible differentiation scenarios

	Homogeneous	Heterogeneous
General-purpose	• Natural monopoly expected • Extended wars of attrition (major prize) fought across price and innovation. • Low contestability resulting in prices approaching monopoly levels • Innovation may slow once monopoly is established due to low contestability • Extensive scope for up- and downstream vertical foreclosure	• Oligopoly expected • Firms compete on unique features, performance, and cost-effectiveness • Differentiation drives prices and profits between monopoly and perfect competition levels – level depends on the degree of substitutability • Incentives for firms to innovate and differentiate as much as possible to maximise competitive edge and market power
Specialised	• Natural monopolies expected, but within segmented specialised sub-markets • Competition for smaller prizes, with wars of attrition fought across price and innovation • Higher contestability due to lower absolute sunk costs, but prices nevertheless approaching monopoly levels • Innovation may slow once monopolies establish, but higher contestability moderates this • Scope for up- and downstream vertical foreclosure, but less so than in the general-purpose/homogeneous scenario	• Series of oligopolies in each specialised market expected: most competitive scenario • Competition centres on innovation and diversity, with models highly optimized for specific use cases • Lower sunk costs lead to frequent entry and exit based on innovation and consumer preferences, pushing prices closer to competitive levels • Intermediate pricing as differentiation drives prices and profits between monopoly and perfect competition levels. Lower barriers to entry, reduced sunk costs, and substitutability further push prices toward competitive levels

Mixed outcomes

How consumer demands for foundation models can and will most efficiently be met remains an open question. The four scenarios outlined above and summarised in Table 2 represent idealised outcomes for analytical purposes. However, it is possible – indeed likely – that the industry is characterised by a combination of two or more of them, a point noted by Vipra & Korinek (2023). Homogeneous general-purpose models may for instance cater to a significant proportion of consumers' needs while niche markets emerge for others which cannot be sufficiently or efficiently served by them.

If this happens, performance and price will be close enough at the boundaries to introduce competition between dissimilar products as multi-dimensional general-purpose models compete for the same customers as specialised models. Even if monopolistic dynamics emerge in some sub-markets (notably ones providing more costly general-purpose models) due to high barriers to entry and homogenisation, the scope for competition between one- and

multi-dimensional products may be a unique characteristic of this industry which introduces competitive dynamics in a distinctive way. Though these will not have the same competitive influence on the general purpose producer's behaviour as true contestability, they can exert sufficient potential pressure so as to be noteworthy. Regulators ought to be aware of this form of competition, protecting it from anticompetitive behaviour.

V. Regulatory implications of natural monopoly dynamics in foundation model production

We now consider potential policy responses to address or pre-empt market failures associated with the concentrating dynamics of foundation model production. Three key findings emerge from our industry analysis. First, foundation model production exhibits subadditive cost functions due to high sunk costs, leading to extensive economies of scale. Second, dynamic analysis shows that initial intense competition is highly probable despite, or because of, this. This competition, while potentially productively inefficient, may be dynamically efficient and welfare-enhancing by incentivizing innovation. Finally, the uncertain trajectory of product differentiation suggests that different competitive outcomes may arise, ranging from a single natural monopoly to sub-market oligopolies. Given the unpredictable nature of future technological developments and consumer demands, predicting which competitive scenario or combination thereof will prevail is impossible. Emerging regulation should therefore be flexible enough to be applicable and effective to all outlined scenarios.

Regulatory approaches must be adaptable to these potential outcomes, ensuring they can accommodate a wide range of industry developments. Effective regulation in this industry must balance the need for oversight in markets with natural monopoly dynamics with the flexibility to adapt to a rapidly changing technological landscape. But regulation is necessary given the potential value of the prizes at stake and the possibility of entrenched monopolies emerging. There are therefore good reasons to regulate this industry pre-emptively, avoiding some of the mistakes made in the digital platform era (Narechania & Sitaraman, 2023b).

Competitive harms and inefficiencies in foundation model production can be divided in two. First, there is the question of who will be the natural monopolist or oligopolist. Given the

intensity of competition and the value of the prize or prizes at stake, there are strong incentives for competing firms to engage in anticompetitive behaviour in their efforts to gain the monopoly position. This is magnified by the downside of losing these wars of attrition given the value of irredeemable sunk costs that are forfeited. Second, however, is the question of how regulators should respond if or when a natural monopoly or oligopoly emerges.

Crucially, monopoly itself is not inherently undesirable. We must differentiate between monopolies that arise due to efficiency and those that do not. As Posner (1969: p.564) notes, "the effort of a businessman to monopolize a market by producing at a cost so low as to drive out his competitors and deter new entry or, the monopoly achieved, to improve his return by lowering his costs still further is not at all reprehensible". Efficient monopolies, or true natural monopolies, require regulation to prevent ills that arise when competition is insufficient to discipline the monopolist. In these cases, the focus should be on addressing the market failures that accompany monopoly rather than viewing monopoly itself as a market failure.

A. Regulating competition for foundation model markets

Natural monopoly is not a market failure but the intensity of competition for promised natural monopoly rents can easily incentivise anticompetitive behaviour given the allure of the prize. Though wars of attrition should sort firms (see Part IV.C), with the most efficient one emerging as the last player standing, a range of anticompetitive behaviours can interfere with that process, resulting in undesirable and inefficient winners. Thus, while monopoly or oligopoly outcomes are not necessarily indicative of market failure, they become so if the wrong firm or firms gain those positions.

Regulation addressing anticompetitive conduct can take two forms: ex-post and ex-ante. Ex-post regulation, which includes traditional competition law, punishes anticompetitive conduct after it has occurred, seeking to address its consequences through retrospective court-ordered remedies (Bostoen & van Wamel, 2023). It is flexible, making it well-suited for new industries, but its effectiveness primarily relies on its deterrent threat. Ex-ante regulation, conversely, seeks to prevent harmful conduct from occurring in the first place by setting out clear and specific banned behaviours or by imposing positive obligations on producers or monopolists (Narechania & Sitaraman, 2023b; Ottaviani & Wickelgren, 2011). This requires

legislators to have identified the sector-specific dynamics and anticompetitive behaviours that cause concentration. Though not applicable to foundation models themselves, the DMA is a leading example of ex-ante regulation.

Competitive harms and traditional ex-post regulation

Natural monopolies are typically achieved through competitive processes. Ex-post regulation exists to protect these. Its strength lies in its flexibility. Its rules and principles which target abuses of dominance, foreclosure, exclusive dealing, tying and bundling, and refusals to deal are sufficiently broad and conduct-oriented to apply to any industry or market (Motta, 2004). It therefore does not need amendment to be applied to the foundation model industry, just as it did not for it to be applied to platform markets (see e.g. *Google Shopping*, 2021; *Microsoft v Commission*, 2007). Accepted practice must evolve, however, with regulators needing to pay careful attention to identify the anticompetitive implications of new forms of conduct early. In the digital platform case, for instance, the use of most-favoured nation clauses or specific types of bundling only came under the scrutiny of competition authorities late.

Though vertical concerns deserve significant attention (e.g. as underlined by Hoppner & Streatfeild, 2023; CMA, 2024a), one practice which has received insufficient consideration in the foundation model context is predatory pricing (Wansley & Weinstein, 2023). Fumagalli & Motta (2013) demonstrate how scale economies like those present in the foundation model market create an incentive for incumbents or first movers to prey on more efficient rivals by pricing goods below marginal cost. Unlike two-sided markets (such as search engines and social media platforms) which can justify low prices for search consumers on the basis that the product they in fact sell is advertising, foundation model production is a one-sided market. The provision of models for free can only be achieved at a loss. Competition authorities may therefore want to consider whether and to what extent firms engage in competitively harmful predatory pricing strategies (Fumagalli et al., 2018). This is a complex question deserving of further analysis, not least because it involves considering the costs, benefits, and incentives behind open-source model provision which, though by definition priced below marginal cost, may be deserving of an exemption.

In terms of vertical concerns, three types of exclusionary practice are likely: exclusive dealing, tying and bundling, and vertical foreclosure. Exclusive dealing, specifically with

input suppliers, is another form of likely anticompetitive behaviour. Though scholarly attention has predominantly focused on the more common practice of downstream exclusive dealing (Fumagalli et al., 2018; Fumagalli & Motta, 2006), it can be practiced in the other direction, with suppliers. This is unlikely to occur in the context of compute, as that industry is more standardised and has no incentive to restrict its customers. This is not the case with data, which for many firms and industries such as journalism is a byproduct of their main product offering and source of income. There is therefore a greater chance that firms will successfully seek to secure exclusive deals for data, at least while the market for that remains young and suppliers don't consider each other to be competitors (Fumagalli & Motta, 2006). A key source of competitive advantage, restricting competitors' access to data through exclusive agreements limits their ability to innovate and compete effectively. Careful attention to such practices is thus warranted.

The potential for input foreclosure by vertically integrated firms producing foundation models has been discussed extensively, including by competition authorities (see e.g. Aut. conc., 2024; Belfield & Anonymous, 2023; Narechania & Sitaraman, 2023; CMA, 2023a, 2024a). As these contributions suggest, there is scope for vertically integrated firms, especially those with monopoly or oligopoly positions in upstream markets for data and computing and over downstream distribution platforms, to leverage their positions to their own competitive advantage. A similar concern is true for firms producing foundation models and cloud computing services as they may seek to bundle foundation models with cheaper inference computing. Ensuring these behaviours do not prevent more efficient foundation model producers from competing is of paramount importance.

In fast-moving markets, ex-post regulation's primary use is in its deterrent effect, which is influenced by the perceived probability of punishment, the extent of expected punishment, and the reputation of the enforcing authority (Dierx et al., 2023). Enforcers have historically been resource-limited, and their reputations are recovering from decades of weak enforcement following the dominance of Chicago School economic thought (Ezrachi & Stucke, 2017a, 2017b; Shapiro, 2020). Cases are generally long, with the *Google Shopping* case infamously only finally being decided after twelve years, by which point the anticompetitive behaviour in question had already caused irreversible competitive harm (*Google and Alphabet v Commission (Google Shopping)*, 2021; *Google Search (Shopping)*, 2017; Marsden, 2020). This reputation failure arguably set the scene for incumbent and

aspiring digital monopolists to engage extensively in anticompetitive behaviours in the 2000s and 2010s.

However, competition authorities however experienced a resurgence in the 2010s and 2020s, especially in the digital sphere, with a significant increase in investigations and cases, especially concerning large technology firms (Lasarte, 2023). In contrast with the platform era, competition authorities have signalled a strong willingness to be active in foundation model markets from the outset, suggesting an explicit intention to amplify this deterrent effect. The CMA and its French counterpart published extensive policy reports in 2023 and 2024 (Aut. conc., 2024; CMA, 2023a, 2024a). The CMA and the US Federal Trade Commission (FTC) initiated enquiries and investigation procedures into several partnerships and mergers in the industry (CMA, 2023b, 2024b; FTC, 2024). The strong desire to signal their intention to deter, prevent and punish anticompetitive conduct in this industry appears to have led the EU Commission, the FTC, the UK CMA, and the US Department of Justice to release a highly unusual joint statement on competition in foundation models in July 2024 (European Commission et al., 2024).

Expedited ex-post regulation

Though the flexibility and universality of ex-post competition law is a strength, past experiences suggest deterrence is not always sufficient. This is particularly true in the context of natural monopoly positions whose long-term attainment firms may consider worth the risk of punishment. Sector-specific ex-post regulatory regimes which prescribe more specific sector-focused rules seek to address this issue by expediting enforcement processes.

Though regulations such as the DMA and the United Kingdom's Digital Markets, Competition and Consumers Act 2024 are often portrayed as ex-ante regulations, they are better understood as providing for a combination of ex-ante rules and expedited ex-post enforcement mechanisms (DMCCA, 2024; DMA, 2022). Both establish prohibitions on anticompetitive behaviours that would also fall foul of competition law (Colangelo, 2022), but give greater investigatory and enforcement powers to regulators to accelerate the process of identifying, punishing, and correcting such behaviours. This is important in the context of possible natural monopolies because it increases the chance of remedying behaviours in the competitive phase of the process before monopolists become entrenched.

DMA-style regulation works in part by creating a regime with a catalogue of specific prohibitions for undertakings (entities engaged in economic activities) identified as having a strategic market status in their industries. In the DMA's case, these include prohibitions on most favoured nation clauses (Art.5(b) DMA; *HRS-Hotel Reservation Service Robert Ragge GmbH*, 2013), self-preferencing (Art.6(1)(d); *Google Shopping*, 2021), and tying and bundling (Art.5(e), Art.5(f), Art.6(1)(b), Art.6(1)(e); *Microsoft v Commission*, 2007). But prohibiting actions by identified undertakings is an act most appropriate for already concentrated markets, not ones that we believe will become concentrated in the long run. Even if the DMA does, as Andriychuk (2023) argues, seek to shape the market, it only does so by creating a special regime to discipline incumbents. This is appropriate and important when it comes to regulating monopolies when they have emerged, as is the case in network-effect-influenced platform markets, but it is ineffective at protecting the process by which they many emerge in the first place.

However, because sources of anticompetitive advantage in foundation model production are predominantly vertical in nature (CMA, 2024a), upstream industries that supply key inputs, notably data and compute, are of significant interest (Carugati, 2024). Protecting this aspect of the competitive process in foundation model production is therefore better served by regimes focused on cloud computing as a sector with strategic market status (Benzina, 2019). Importantly, cloud computing services fall within the scope of both the DMA (Art.2(2)(i) DMA, 2022) and the DMCCA.

Complementary ex-ante regulation

While prohibiting anti-competitive behaviour addresses market issues reactively, ex-ante obligations introduce proactive measures that impose positive duties on firms. These measures predominantly target behaviours that create artificial barriers to entry and reduce contestability, by seeking to prohibit or pre-empt them. Because foundation model monopoly dynamics and competitive advantages are derived from the supply side rather than demand-side network effects, the main scope for ex-ante regulation focuses on supply, and therefore inputs. In this context, these inputs primarily include data and cloud computing services.

Ex-ante obligations can either artificially increase competition or enhance ex-post enforcement through transparency and reporting requirements (e.g., Arts. 11, 14, and 15

DMA). In the context of foundation models and cloud computing services, regulators should ensure that such requirements enable them to prevent integrated firms from engaging in strategies such as margin pricing. Although the DMA does not explicitly set FRAND (Fair, Reasonable, and Non-Discriminatory) standards, provisions like Article 6(11) and Article 6(12) effectively set FRAND thresholds for fair access to named services. This standard could be usefully extended to cloud computing service provision for foundation model producers.

Interoperability (Art. 6(7) DMA) and data portability (Art. 12) obligations foster competition by allowing easier market entry and data transfer between services, and by preventing firms from introducing unreasonable switching costs for consumers. While these should be balanced by considerations of necessity and proportionality, especially with regard to innovation, they are essential to prevent vendor lock-in and ensure level playing fields by allowing developers to deploy and test foundation models across cloud service platforms and environments without being restricted to a single provider's ecosystem. Equally, maximising interoperability in chip architectures ensures that foundation models can leverage a broad range of hardware resources, enhancing performance and scalability. Although there may be legitimate reasons to develop specific chip architectures for specific model architectures, promoting the adoption of open standards for interoperability and compatibility can facilitate competition and prevent vendor lock-in. Though switching costs may be high for foundation model customers as they need to go through the fine-tuning process each time they change supplier, these are not artificially imposed. Given current technologies, it is not probable that interoperability requirements should be imposed on foundation model producers.

B. Regulating foundation model natural monopolies

Contestability

As Baumol et al. (1982) demonstrate, the more contestable a market is, the greater the pressure and incentive for incumbent monopolists to behave in a welfare-maximising manner. The higher the potential for competition, the lower the need for regulation to surrogate for competition and artificially correct its inefficiencies. If sunk costs are unavoidable sources of barriers to entry in the foundation model industry, other barriers to entry not inherent to model production are not. As such, the matters and measures discussed in Section V.A which address competition for foundation model markets remain significant when those markets

have concentrated or monopolised. As we have established however, the foundation model market, even absent anticompetitive behaviour, is far from contestable. Other market regulation is therefore necessary to prevent the market failures likely to ensue from the natural monopoly or oligopoly outcomes we have identified as highly likely in this industry.

A role for ex-ante or utility-style regulation?

Concentration or monopolisation having occurred, new possible market failures and sources of competitive harm emerge. Apart from the typical monopoly pricing problem, foundation model production may go from being the subject of exclusionary practices to their source as foundation model monopolists or oligopolists compete with downstream firms seeking to use their model to develop fine-tuned versions or applications incorporating it. Foundation models involve complex technologies and specific operational practices requiring distinct regulatory approaches compared to the DMA which primarily targets digital platforms.

Monopoly prices may have to be accepted as the price to pay for this innovation, both because of Schumpeterian arguments and because of the unfeasibility of pricing regulations. Traditional natural monopoly utility-style price regulations depend on the regulator having knowledge of the firm's cost function (Liston, 1993; Vogelsang, 2002). Franchise bidding meanwhile is problematic if the market demands differentiation. It also disincentivises innovation once a franchise is won (Williamson, 1976). Minimising monopoly price distortions by maximising contestability is the best possible outcome, especially if the heterogeneous or mixed scenarios outlined in Section IV.D play out, in which case the availability of substitute goods should exert downward pressure on prices.

Ex-ante regulation may be necessary to prevent foundation monopolists and oligopolists from using their centralised power to restrict, control, or adversely influence downstream uses and implementations of their models to anticompetitive ends. If monopolists develop their own downstream applications and fine-tuned versions of foundation models, there are important incentives for abuse of dominance including restricting the extent to which downstream competitors can interact with the model and fine-tune it (e.g. access only possible via restrictive APIs). Equally, concentration at the foundation model level facilitates artificially generated monopsonistic power if foundation model monopolists or oligopolists restrict or determine which cloud computing service provider their models can be run on. Vertically integrated firms which produce a monopolistic foundation model and provides cloud

computing services will have particularly good incentives to engage in such behaviour. As such, specific ex-ante regulation may be necessary to mandate cloud service interoperability for foundation models and to limit the extent to which concentration at the foundation model production level is leveraged to achieve competitive advantages at the downstream application level of production.

Conclusion

The economic analysis of the foundation model industry conducted in the first part of this paper has given us a number of key insights which can help inform regulatory approaches seeking to ensure competitive harms do not occur in it. The cost functions of foundation model production are shown to be subadditive given the extreme ratio between sunk fixed costs and marginal costs. Despite the absence of network effects, strong natural monopoly dynamics therefore characterise the industry. At the time of writing, however, no monopolist has emerged to take control of foundation model markets, largely because the technology's novelty and the rapid pace of innovation combined with the expected value of the monopolist's prize means wars of attrition have developed. In the short run, this is highly beneficial to consumers as they benefit from the innovation and prices remain at or close to marginal cost. Over the long run, however, we can expect natural monopoly dynamics to exert important influence on the foundation model industry.

But while foundation model cost functions are subadditive, the uncertainty surrounding the nature of the products that will ultimately emerge and be provided by the industry means we can only speculate as to precisely how this will influence concentration. First, there is the question of specialisation: it is not yet clear whether the industry will consolidate around general purpose models, or if the technological limitations of what can be achieved by general models and market demand will instead lead to consumer needs being best fulfilled by niche models specialising in certain tasks and/or fields. Second, is the question of homogeneity: it is not clear whether either general purpose or specialised models will converge in performance and features, or if there will be scope for significant enough differentiation so as to introduce a non-price competitive influence. A taxonomy was

developed to demonstrate the diversity of potential foundation model industry futures and to outline how competition might, or might not, materialise in each.

The inescapable unpredictability of this industry's technological future combined with the clear presence of natural monopoly dynamics, sets down a challenge when seeking to protect the competitive process and avoid harmful market failures. First and foremost, it is necessary to recognise that the emergence of natural monopolies at the foundation model level of production is not inherently bad. So long as it emerges naturally, monopoly can be the most efficient way of delivering the product. Regulation therefore should not seek to prevent the emergence of foundation model monopolies or oligopolies. Instead, monopolies should be acknowledged and accepted, and regulation should pursue two aims: (1) ensuring the monopolies that emerge do so for natural reasons by effectively regulating and deterring anticompetitive conduct through ex-post and ex-ante regulation, and (2) carefully surrogating for competition with limited targeted interventions, disciplining monopolists where the market can no longer do so itself. This is challenging given the industry's pervasively uncertain future demonstrated in Part VI, but it is achievable through efficient ex-post enforcement and principled ex-ante contestability-maximising regulation.

Bibliography

Acemoglu, D., & Johnson, S. (2023, June 10). Opinion | Big Tech Is Bad. Big A.I. Will Be Worse. *The New York Times*. https://www.nytimes.com/2023/06/09/opinion/ai-big-tech-microsoft-google-duopoly.html?searchResultPosition=1

Altman, S. (2023, February 14). Planning for AGI and beyond. *OpenAI*. https://openai.com/index/planning-for-agi-and-beyond/

Andriychuk, O. (2023). Do DMA obligations for gatekeepers create entitlements for business users? *Journal of Antitrust Enforcement*, *11*(1), 123–132. https://doi.org/10.1093/jaenfo/jnac034

Areeda, P., & Turner, D. F. (1975). Predatory Pricing and Related Practices under Section 2 of the Sherman Act. *Harvard Law Review*, *88*(4), 697–733. https://doi.org/10.2307/1340237

Autorité de la concurrence. (2024). *AVIS 24-A-05 du 28 juin 2024 relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative*. Autorité de la concurrence de la République Française. https://www.autoritedelaconcurrence.fr/sites/default/files/integral_texts/2024-07/24a05_merged.pdf

Bain, J. S. (1949). A Note on Pricing in Monopoly and Oligopoly. *The American Economic Review*, *39*(2), 448–464.

Barr, W. P. (2024, May 27). Opinion | Big Tech's Budding AI Monopoly. *Wall Street Journal*. https://www.wsj.com/articles/big-techs-budding-ai-monopoly-40280c15

Baumol, W. J. (1977). On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry. *The American Economic Review*, *67*(5), 809–822.

Baumol, W. J., Bailey, E. E., & Willig, R. D. (1977). Weak Invisible Hand Theorems on the Sustainability of Multiproduct Natural Monopoly. *The American Economic Review*, *67*(3), 350–365.

Baumol, W. J., Panzar, J. C., & Willig, R. D. (1982). Contestable markets and the theory of industry structure. In *Contestable markets and the theory of industry structure*. Harcourt Brace Jovanovich.

Beckert, J. (2016). *Imagined Futures: Fictional Expectations and Capitalist Dynamics*. Harvard University Press.

Belfield, H. & Anonymous. (2023). Compute and Antitrust: Regulatory implications of the AI hardware supply chain, from chip design to foundation model APIs. *Proceedings of the 40th International Conference on Machine Learning*.

Benzina, K. (2019). Cloud Infrastructure-as-a-Service as an Essential Facility: Market Structure, Competition, and the Need for Industry and Regulatory Solutions. *Berkeley Technology Law Journal*, *34*(1), 119–142.

Bindley, K. (2024, March 27). The Fight for AI Talent: Pay Million-Dollar Packages and Buy Whole Teams. *The Wall Street Journal*. https://www.wsj.com/tech/ai/the-fight-for-ai-talent-pay-million-dollar-packages-and-buy-whole-teams-c370de2b

Blaug, M. (2001). Is Competition Such a Good Thing? Static Efficiency versus Dynamic Efficiency. *Review of Industrial Organization*, *19*(1), 37–48. https://doi.org/10.1023/A:1011160622792

Blut, M., Wang, C., Wünderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, *49*(4), 632–658. https://doi.org/10.1007/s11747-020-00762-y

Bobrow, D. G., Mittal, S., & Stefik, M. J. (1986). Expert systems: Perils and promise. *Communications of the ACM*, *29*(9), 880–894. https://doi.org/10.1145/6592.6597

Bommasani, R., & Liang, P. (2021, October 18). *Reflections on Foundation Models*. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/news/reflections-foundation-models

Bosley, M., Jacobs-Harukawa, M., Licht, H., & Hoyle, A. (2023). *Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research*.

Bostoen, F., & van Wamel, D. (2023). Antitrust Remedies: From Caution to Creativity. *Journal of European Competition Law & Practice*, *14*(8), 540–552. https://doi.org/10.1093/jeclap/lpad051

Brock, W. A. (1983). Contestable Markets and the Theory of Industry Structure: A Review Article. *Journal of Political Economy*, *91*(6), 1055–1066.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Bulow, J., & Klemperer, P. (1999). The Generalized War of Attrition. *American Economic Review*, *89*(1), 175–189. https://doi.org/10.1257/aer.89.1.175

Carugati, C. (2024). *The Competitive Relationship Between Cloud Computing and Generative AI*. Bruegel. https://www.ssrn.com/abstract=4738738

Center for Research on Foundation Models (CRFM). (n.d.). *Holistic Evaluation of Language Models (HELM)*. Retrieved 4 August 2024, from https://crfm.stanford.edu/helm/classic/latest/#/leaderboard

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017). *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs* (arXiv:1606.00915). arXiv. https://doi.org/10.48550/arXiv.1606.00915

Colangelo, G. (2022). The European Digital Markets Act and antitrust enforcement: A liaison dangereuse. *European Law Review*, *47(5)*(5), 597–621.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). *The rising costs of training frontier AI models* (arXiv:2405.21015). arXiv. http://arxiv.org/abs/2405.21015

Crane, D. A. (2022). Antitrust as an Instrument of Democracy. *Duke Law Journal*, *72*(21), 21–40.

Cui, H., Wang, C., Maan, H., & Wang, B. (2023). *scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI* (p. 2023.04.30.538439). bioRxiv. https://doi.org/10.1101/2023.04.30.538439

David, P. A., & Greenstein, S. (1990). The Economics Of Compatibility Standards: An Introduction To Recent Research. *Economics of Innovation and New Technology*, *1*(1–2), 3–41. https://doi.org/10.1080/10438599000000002

Dierx, A., Ilzkovitz, F., Pataracchia, B., & Pericoli, F. (2023). Modelling the Diffusion of the Deterrent Effects of Competition Policy. *Journal of Competition Law & Economics*, *19*(2), 277–311. https://doi.org/10.1093/joclec/nhad004

Digital Markets, Competition and Consumers Act, § c.13 (2024). https://www.legislation.gov.uk/ukpga/2024/13/enacted

Ducci, F. (2022). Natural monopolies in digital platform markets. In *Natural monopolies in digital platform markets*. Cambridge University Press.

Edgeworth, F. Y. (1911). Contributions to the Theory of Railway Rates. *The Economic Journal*, *21*(83), 346–370. https://doi.org/10.2307/2222325

Efrati, A., & Holmes, A. (2024, July 24). Why OpenAI Could Lose $5 Billion This Year. *The Information*. https://www.theinformation.com/articles/why-openai-could-lose-5-billion-this-year

Ely, R. T. (1894). Natural Monopolies and the Workingman. A Programme of Social Reform. *The North American Review*, *158*(448), 294–303.

European Commission, UK Competition and Markets Authority, US Department of Justice, & US Federal Trade Commission. (2024, July 23). *Joint statement on competition in generative AI foundation models and AI products*. https://competition-policy.ec.europa.eu/document/download/79948846-4605-4c3a-94a6-044e344acc33_en?filename=20240723_competition_in_generative_AI_joint_statement_COMP-CMA-DOJ-FTC.pdf

Ezrachi, A., & Stucke, M. E. (2017a). *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy*. Harvard University Press. https://doi.org/10.4159/9780674973336

Ezrachi, A., & Stucke, M. E. (2017b, December 15). The Rise, Fall, and Rebirth of the U.S. Antitrust Movement. *Harvard Business Review*. https://hbr.org/2017/12/the-rise-fall-and-rebirth-of-the-u-s-antitrust-movement

Faraboschi, P., Giles, E., Hotard, J., Owczarek, K., & Wheeler, A. (2024). *Reducing the Barriers to Entry for Foundation Model Training* (arXiv:2404.08811). arXiv. http://arxiv.org/abs/2404.08811

Floridi, L. (2020). AI and Its New Winter: From Myths to Realities. *Philosophy & Technology*, *33*(1), 1–3. https://doi.org/10.1007/s13347-020-00396-6

Fortuna, P., & Gorbaniuk, O. (2022). What Is Behind the Buzzword for Experts and Laymen: Representation of "Artificial Intelligence" in the IT-Professionals' and Non-Professionals' Minds. *Europe's Journal of Psychology*, *18*(2), 207–218. https://doi.org/10.5964/ejop.5473

Fudenberg, D., & Tirole, J. (1984). The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look. *The American Economic Review*, *74*(2), 361–366.

Fumagalli, C., & Motta, M. (2006). Exclusive Dealing and Entry, when Buyers Compete. *American Economic Review*, *96*(3), 785–795. https://doi.org/10.1257/aer.96.3.785

Fumagalli, C., & Motta, M. (2013). A Simple Theory of Predation. *The Journal of Law & Economics*, *56*(3), 595–631. https://doi.org/10.1086/672951

Fumagalli, C., Motta, M., & Calcagno, C. (2018). *Exclusionary Practices: The Economics of Monopolisation and Abuse of Dominance*. Cambridge University Press. https://doi.org/10.1017/9781139084130

Gans, J. S. (2024). *Market Power in Artificial Intelligence* (Working Paper 32270). National Bureau of Economic Research. https://doi.org/10.3386/w32270

Geroski, P. A. (2003). Competition in Markets and Competition for Markets. *Journal of Industry, Competition and Trade*, *3*(3), 151–166. https://doi.org/10.1023/A:1027457020332

Gilbert, A. (2024, March 7). Google-Reddit AI Deal Heralds New Era in Social Media Licensing. *Bloomberg Law*. https://news.bloomberglaw.com/ip-law/google-reddit-ai-deal-just-the-start-for-social-media-licensing

Google and Alphabet v Commission (Google Shopping), T-612/17, EU:T:2021:763 (CJEU 10 November 2021). https://curia.europa.eu/juris/document/document.jsf?text=&docid=250881&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=6748881

Google Search (Shopping), AT.39740 (European Commission 27 June 2017). https://ec.europa.eu/competition/antitrust/cases/dec_docs/39740/39740_14996_3.pdf

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). *Textbooks Are All You Need* (arXiv:2306.11644). arXiv. https://doi.org/10.48550/arXiv.2306.11644

Hagiu, A., & Wright, J. (2023). Data-enabled learning, network effects, and competitive advantage. *The RAND Journal of Economics*, *54*(4), 638–667. https://doi.org/10.1111/1756-2171.12453

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Hille, K., & Liu, Q. (2023, August 23). *Supply chain shortages delay tech sector's AI bonanza*. https://www.ft.com/content/c7e9cfa9-3f68-47d3-92fc-7cf85bcb73b3

Hirsch-Kreinsen, H. (2023). Artificial intelligence: A "promising technology". *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01629-w

Hobbhahn, M., Heim, L., & Aydos, G. (2023, November 9). *Trends in Machine Learning Hardware*. Epoch AI. https://epochai.org/blog/trends-in-machine-learning-hardware

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. de L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022). *Training Compute-Optimal Large Language Models* (arXiv:2203.15556). arXiv. http://arxiv.org/abs/2203.15556

Hoppner, T., & Streatfeild, L. (2023). *ChatGPT, Bard & Co.: An Introduction to AI for Competition and Regulatory Lawyers* (SSRN Scholarly Paper 4371681). https://doi.org/10.2139/ssrn.4371681

HRS-Hotel Reservation Service Robert Ragge GmbH, B.9 — 66/10 (BKartA 20 December 2013).

Joskow, P. L. (2007). Chapter 16: Regulation of Natural Monopoly. In A. M. Polinsky & S. Shavell (Eds.), *Handbook of Law and Economics* (Vol. 2, pp. 1227–1348). Elsevier. https://doi.org/10.1016/S1574-0730(07)02016-6

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (arXiv:2001.08361). arXiv. http://arxiv.org/abs/2001.08361

Katz, M. L., & Shapiro, C. (1985). Network Externalities, Competition, and Compatibility. *The American Economic Review*, *75*(3), 424–440.

Katz, M. L., & Shapiro, C. (1994). Systems Competition and Network Effects. *Journal of Economic Perspectives*, *8*(2), 93–115. https://doi.org/10.1257/jep.8.2.93

Kautz, H. (2022). The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, *43*(1), Article 1. https://doi.org/10.1002/aaai.12036

Kelleher, J. D. (2019). *Deep Learning* (1st ed.). MIT Press. https://doi.org/10.7551/mitpress/11171.001.0001

Knight, F. H. (1921). Cost of Production and Price over Long and Short Periods. *Journal of Political Economy*, *29*(4), 304–335. https://doi.org/10.1086/253349

Kucharavy, D., & De Guio, R. (2011). Application of S-shaped curves. *Procedia Engineering*, *9*, 559–572. https://doi.org/10.1016/j.proeng.2011.03.142

Küspert, S., Moës, N., & Dunlop, C. (2023, February 10). The value chain of general-purpose AI. *Ada Lovelace Institute*. https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/

Lasarte, D. (2023, January 24). *The ongoing big tech antitrust cases to watch in 2023*. Quartz. https://qz.com/antitrust-cases-big-tech-2023-guide-1849995493

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lehdonvirta, V. (2022). *Cloud empires: How digital platforms are overtaking the state and how we can regain control*. The MIT Press.

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., & Lee, Y. T. (2023). *Textbooks Are All You Need II: Phi-1.5 technical report* (arXiv:2309.05463). arXiv. https://doi.org/10.48550/arXiv.2309.05463

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … Koreeda, Y. (2023). *Holistic Evaluation of Language Models* (arXiv:2211.09110). arXiv. http://arxiv.org/abs/2211.09110

Liston, C. (1993). Price-cap versus rate-of-return regulation. *Journal of Regulatory Economics*, *5*(1), 25–48. https://doi.org/10.1007/BF01066312

Ma, Y., Wang, Z., Yang, H., & Yang, L. (2020). Artificial intelligence applications in the development of autonomous vehicles: A survey. *IEEE/CAA Journal of Automatica Sinica*, *7*(2), 315–329. IEEE/CAA Journal of Automatica Sinica. https://doi.org/10.1109/JAS.2020.1003021

Marar, S. (2024). *Artificial Intelligence and Antitrust Law: A Primer* (Mercatus Special Study). Mercatus Center at George Mason University. https://www.ssrn.com/abstract=4745321

Marsden, P. (2020). Google Shopping for the Empress's New Clothes –When a Remedy Isn't a Remedy (and How to Fix it). *Journal of European Competition Law & Practice*, *11*(10), 553–560. https://doi.org/10.1093/jeclap/lpaa050

Marshall, A. (1898). Distribution and Exchange. *The Economic Journal*, *8*(29), 37–59. https://doi.org/10.2307/2956696

Mauran, C. (2024, June 21). *All the media companies that have licensing deals with OpenAI (so far)*. Mashable. https://mashable.com/article/all-the-media-companies-that-have-licensing-deals-with-openai-so-far

Microsoft v Commission, T-201/04, EU:T:2007:289 (CJEU 17 September 2007). https://curia.europa.eu/juris/showPdf.jsf?text=&docid=62940&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=7429738

Milgrom, P., & Roberts, J. (1982). Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis. *Econometrica*, *50*(2), 443–459. https://doi.org/10.2307/1912637

Mill, J. S. (1848). *Principles of political economy with some of their applications to social philosophy*. J.W. Parker.

Mitra, A., Khanpour, H., Rosset, C., & Awadallah, A. (2024). *Orca-Math: Unlocking the potential of SLMs in Grade School Math* (arXiv:2402.14830). arXiv. https://doi.org/10.48550/arXiv.2402.14830

Mosca, M. (2008). On the origins of the concept of natural monopoly: Economies of scale and competition. *The European Journal of the History of Economic Thought*, *15*(2), 317–353. https://doi.org/10.1080/09672560802037623

Motta, M. (2004). *Competition Policy: Theory and Practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804038

Narechania, T. N. (2021). Machine Learning as Natural Monopoly. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3810366

Narechania, T. N., & Sitaraman, G. (2023a). An Antimonopoly Approach to Governing Artificial Intelligence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4597080

Narechania, T. N., & Sitaraman, G. (2023b). An Antimonopoly Approach to Governing Artificial Intelligence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4597080

Naughton, J. (2024, April 20). The big tech firms want an AI monopoly – but the UK watchdog can bring them to heel. *The Observer*. https://www.theguardian.com/commentisfree/2024/apr/20/the-big-tech-firms-want-an-ai-monopoly-but-the-uk-watchdog-can-bring-them-to-heel

Newsome, E. C. (2020). Cloudy with a Chance of Monopolization. *Intellectual Property and Technology Law Journal*, *25*(1), 37–62.

Niyato, D., Chaisiri, S., & Lee, B.-S. (2009). Economic analysis of resource market in cloud computing environment. *2009 IEEE Asia-Pacific Services Computing Conference (APSCC)*, 156–162. https://doi.org/10.1109/APSCC.2009.5394127

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., & Horvitz, E. (2023). *Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine* (arXiv:2311.16452). arXiv. https://doi.org/10.48550/arXiv.2311.16452

Ofcom. (2023). *Cloud services market study final report*. https://www.ofcom.org.uk/__data/assets/pdf_file/0027/269127/Cloud-services-market-study-final-report.pdf

Ottaviani, M., & Wickelgren, A. L. (2011). Ex ante or ex post competition policy? A progress report. *International Journal of Industrial Organization*, *29*(3), 356–359. https://doi.org/10.1016/j.ijindorg.2011.02.004

Pareto, V., Montesano, A., Zanni, A., Bruni, L., Chipman, J. S., McLure, M., Pareto, V., Montesano, A., Zanni, A., Bruni, L., Chipman, J. S., & McLure, M. (Eds.). (2020). *Manual of Political Economy: A Critical and Variorum Edition*. Oxford University Press.

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2022). *The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink*.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance*. Free Press.

Posner, R. A. (1969). Natural Monopoly and Its Regulation. *Stanford Law Review*, *21*(3), 548. https://doi.org/10.2307/1227624

Posner, R. A. (1975). The Social Costs of Monopoly and Regulation. *Journal of Political Economy*, *83*(4), 807–827. https://doi.org/10.1086/260357

Rawat, W., & Wang, Z. (2017). Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Computation*, *29*(9), 2352–2449. Neural Computation. https://doi.org/10.1162/neco_a_00990

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) (Text with EEA Relevance), 265 OJ L (2022). http://data.europa.eu/eli/reg/2022/1925/oj/eng

Robertson, D. (2024, June 3). The looming AI monopolies. *POLITICO*. https://www.politico.com/newsletters/digital-future-daily/2024/01/18/the-looming-ai-monopolies-00136400

Rochet, J.-C., & Tirole, J. (2003). Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, *1*(4), 990–1029. https://doi.org/10.1162/154247603322493212

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience*, *11*(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

Schumpeter, J. A. (1944). *Capitalism, socialism, and democracy* (Second Impression). Allen & Unwin.

Shaked, A., & Sutton, J. (1982). Relaxing Price Competition Through Product Differentiation. *The Review of Economic Studies*, *49*(1), 3–13. https://doi.org/10.2307/2297136

Shapiro, C. (2020). Antitrust: What Went Wrong and How to Fix It The Future of Antitrust: Cover Stories: Papers, Comments, and Articles on the Future of Antitrust. *Antitrust*, *35*(3), 33–45.

Sharkey, W. W. (1982). *The theory of natural monopoly*. University Press.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Soliani, G. (2024, February 29). Empowering the digital-first business professional in the foundation model era. *IBM Blog*. https://www.ibm.com/blog/empowering-the-digital-first-business-professional-in-the-foundation-model-era/

Sraffa, P. (1925). Sulle relazioni fra costo e quantità prodotta. *Annali Di Economia*, *2*, 277–328.

Stanford Institute for Human-Centered Artificial Intelligence. (2024). *Artificial Intelligence Index Report 2024*. https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_2024_AI-Index-Report.pdf

Stigler, G. J. (1951). The Division of Labor is Limited by the Extent of the Market. *Journal of Political Economy*, *59*(3), 185–193.

Stigler, G. J. (1957). Perfect Competition, Historically Contemplated. *Journal of Political Economy*, *65*(1), 1–17.

Sutton, J. (1991). *Sunk costs and market structure: Price competition, advertising, and the evolution of concentration*. MIT Press.

Sutton, J. (1996). Technology and market structure. *European Economic Review*, *40*(3), 511–530. https://doi.org/10.1016/0014-2921(95)00065-8

Sutton, J. (1998). *Technology and market structure: Theory and history*. MIT Press.

The New York Times Company v. Microsoft Corporation, OpenAI, Inc., 1:23-cv-11195 (United States District Court for the Southern District of New York 27 December 2023). https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf

Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2021). Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. *IEEE Spectrum*, *58*(10), 50–55. https://doi.org/10.1109/MSPEC.2021.9563954

Tirole, J. (1988). The theory of industrial organization. In *The theory of industrial organization*. MIT Press.

Tsaih, R.-H., Chang, H.-L., Hsu, C.-C., & Yen, D. C. (2023). The AI Tech-Stack Model. *Communications of the ACM*, *66*(3), 69–77. https://doi.org/10.1145/3568026

UK Competition and Markets Authority. (2023a). *AI Foundation Models: Initial Report*. UK Competition and Markets Authority.

UK Competition and Markets Authority. (2023b, December 8). *Microsoft / OpenAI partnership merger inquiry*. GOV.UK. https://www.gov.uk/cma-cases/microsoft-slash-openai-partnership-merger-inquiry

UK Competition and Markets Authority. (2024a). *AI Foundation Models: Technical update report*. https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf

UK Competition and Markets Authority. (2024b, July 16). *Microsoft / Inflection inquiry*. GOV.UK. https://www.gov.uk/cma-cases/microsoft-slash-inflection-ai-inquiry

US Federal Trade Commission. (2024, January 24). *FTC Launches Inquiry into Generative AI Investments and Partnerships*. Federal Trade Commission. https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships

Villalobos, P., & Atkinson, D. (2023). *Trading Off Compute in Training and Inference*. Epoch AI. https://epochai.org/blog/trading-off-compute-in-training-and-inference

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). *Will we run out of data? Limits of LLM scaling based on human-generated data* (arXiv:2211.04325). arXiv. http://arxiv.org/abs/2211.04325

Vipra, J., & Korinek, A. (2023). *Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT*. Brookings. https://www.brookings.edu/wp-content/uploads/2023/09/Market-concentration-implications-of-foundation-models-FINAL-1.pdf

Vogelsang, I. (2002). Incentive Regulation and Competition in Public Utility Markets: A 20-Year Perspective. *Journal of Regulatory Economics*, *22*(1), 5–27.

Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

Wansley, M., & Weinstein, S. (2023). *Venture Predation* (SSRN Scholarly Paper 4437360). https://doi.org/10.2139/ssrn.4437360

Weitzman, M. L. (1983). Contestable Markets: An Uprising in the Theory of Industry Structure: Comment. *American Economic Review*, *73*(3), 486–487.

White, J. M. (2016). Moving applications: A multilayered approach to mobile computing. In *Code and the City*. Routledge.

Williamson, O. (1976). Franchise Bidding for Natural Monopolies—In General and with Respect to CATV. *Bell Journal of Economics*, *7*(1), 73–104.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). *BloombergGPT: A Large Language Model for Finance* (arXiv:2303.17564). arXiv. https://doi.org/10.48550/arXiv.2303.17564