Martens, Bertin

Working Paper

What should be done about Google's quasi-monopoly in search? Mandatory data sharing versus AI-driven technological competition

Bruegel Working Paper, No. 10/2023

Provided in Cooperation with:
Bruegel, Brussels

Suggested Citation: Martens, Bertin (2023) : What should be done about Google's quasi-monopoly in search? Mandatory data sharing versus AI-driven technological competition, Bruegel Working Paper, No. 10/2023, Bruegel, Brussels

This Version is available at:
https://hdl.handle.net/10419/274216

WORKING PAPER | ISSUE 10/2023 | 06 JULY 2023

WHAT SHOULD BE DONE ABOUT GOOGLE'S QUASI-MONOPOLY IN SEARCH? MANDATORY DATA SHARING VERSUS AI-DRIVEN TECHNOLOGICAL COMPETITION

BERTIN MARTENS

The first part of this paper focuses on competition between search engines that match user queries with webpages. User welfare, as measured by click-through rates on top-ranked pages, increases when network effects attract more users and generate economies of scale in data aggregation. However, network effects trigger welfare concerns when a search engine reaches a dominant market position. The EU Digital Markets Act (DMA) imposes asymmetric data sharing obligations on very large search engines to facilitate competition from smaller competitors. We conclude from the available empirical literature on search-engine efficiency that asymmetric data sharing may increase competition but may also reduce scale and user welfare, depending on the slope of the search-data learning curve. We propose policy recommendations to reduce tension between competition and welfare, including (a) symmetric data sharing between all search engines irrespective of size, and (b) facilitate user real-time search history and profile-data portability to competing search engines.

The second part of the paper focuses on the impact of recent generative AI models, such as Large Language Models (LLMs), chatbots and answer engines, on competition in search markets. LLMs are pre-trained on very large text datasets, prior to usage. They do not depend on user-driven network effects. That avoids winner-takes-all markets. However, high fixed algorithmic learning costs and input markets bottlenecks (webpage indexes, copyright-protected data and hyperscale cloud infrastructure) make entry more difficult. LLMs produce semantic responses (rather than web pages) in response to a query. That reduces cognitive processing costs for users but may also increase ex-post uncertainty about the quality of the output. User responses to this trade-off will determine the degree of substitution or complementarity between search and chatbots. We conclude that, under certain conditions, a competitive chatbot markets could crowd out a monopolistic search engine market and may make DMA-style regulatory intervention in search engines redundant.

The paper concludes with some policy recommendations.

Keywords: search engines, answer engines, generative AI, large language models, chatbots, ChatGPT, digital competition policy, data governance, access to data, economies of scale and scope in data aggregation.
JEL codes: K21, D23, D43

Bertin Martens (bertin.martens@bruegel.org) is a Visiting Fellow at Bruegel

Recommended citation:
Martens, B. (2023) 'What should be done about Google's quasi-monopoly in search? Mandatory data sharing versus AI-driven technological competition', *Working Paper* 10/2023, Bruegel

bruegel

1 Introduction

Search engines are a crucial gateway to access online services in modern digital economies. When a single search engine – Google Search – reaches a dominant market position, covering about 90 percent of all searches[1], the lack of competition in search may distort downstream services markets that depend on referrals from search engines. This has already led to several competition cases against Google Search, for example the Google Shopping case in the European Union (Deutscher, 2021), and reports on Google's alleged abuse of dominance (US House of Representatives, 2020; Scott-Morton and Dinielli, 2020). Researchers have attributed Google's dominance to data-driven network effects: more users generate more data, which improves the quality of search and therefore attracts even more users to the search engine (Prüfer and Schotmuller, 2022). The economic impact of network effects is ambiguous. They may increase user welfare through better services but may also reduce user welfare because of reduced competition in downstream services markets.

EU competition policymakers focused on the negative competition effect and concluded that breaking the network-effects feedback loop could solve this. To that end, the EU Digital Markets Act (DMA, 2022) imposes obligations on very large 'gatekeeper' search engines, requiring them to share user query and click data with other smaller search engines. DMA Art 6(11) states that gatekeepers "*shall provide to any third-party undertaking providing online search engines, at its request, with access on fair, reasonable and non-discriminatory terms to ranking, query, click and view data in relation to free and paid search generated by end users on its online search engines*". Giving competing search engines access to data collected by the dominant incumbent will facilitate market entry. Competitors will no longer depend on network effects to accumulate the necessary user data to run an efficient search engine. That should eliminate monopolistic search market problems.

However, policymakers may face tension between competition policy and user welfare objectives. 'Classic' competition policy seeks to maximise consumer welfare by promoting competition in markets, avoiding or suppressing the emergence of dominant players with large market shares and replacing them with many competing players with smaller market shares. This competition policy objective should remain valid in the digital platform economy (Digital Competition Expert Panel, 2019). It is reflected in the stated policy objective of the DMA: "*to ensure contestability and fairness for the markets in the digital sector*" (DMA, Recital 7).

In the case of search engines, this view would run into problems if more competition between search engines decreased the quality of search and user satisfaction with search results, because competition fragments user data across many search engines. Our first research question in this paper is to examine how likely these negative effects on consumer welfare are. We do this by reviewing several recent empirical papers on search-engine efficiency. They reveal the importance of user data collection, especially for rare keywords that represent a majority of all searches, and confirm that smaller market shares reduce search engine access to rare keyword data. We then explore whether an appropriate design of search-data-sharing governance mechanisms could reduce the tension between competition and welfare. We conclude that symmetric search-data sharing between all search engines, rather than the asymmetric sharing mechanism foreseen in the DMA, may achieve this.

Apart from data sharing, there may be other ways to instil competition in search markets. In the second part of the paper, we explore a second research question: can recent substantial changes in search-engine technology lead to more competition, and do they make the DMA search-data-sharing obligations redundant? The recent arrival of generative AI with Large Language Models (LLMs) and

[1] Source: Similarweb, March 2023, https://www.similarweb.com/engines/.

chatbots or answer engines, such as Chat-GPT[2], and their ability to produce a more elaborate and in-depth semantic reply to user queries, compared to search engines, represents a substantial technological innovation and a natural experiment to detect the impact of technology on search-market positions. Rather than ranking webpages and letting users search for the desired answer in these pages, LLMs give users a reasoned natural language answer to a query. Hence the label 'answer engines', as distinct from 'search engines'. Answer engine LLMs are pre-trained on very large text databases, prior to usage. Unlike search engines, they do not rely on user network effects to climb a learning curve with increasing market share. On the one hand, the absence of network effects may facilitate market entry for smaller players. On the other hand, LLMs require access to oligopolistic input markets, including a global index of web pages, an inventory that only Google Search and Microsoft Bing have compiled, and hyperscale computing infrastructure. Moreover, pre-training LLM models comes at a high fixed cost for new market entrants. The net effect of these two opposing forces on market entry remains an open question.

In practice, search and answer engines are partial substitutes. For semantically simple queries, consumers may prefer search. For semantically complex queries, answer engines reduce user transaction costs because they do the semantic processing that search engines cannot do (Wu *et al*, 2020). At the same time, answer engines may not be entirely reliable and produce poor quality or even erroneous responses. They may 'hallucinate'. Producers of search services may combine search and answer engines in a single service to capture users that prefer one or the other. That positions search and answer engines as at least partial substitutes, depending on user preferences regarding the trade-off between semantic transaction costs and *ex-post* uncertainty about the outcome. We conclude that input market bottlenecks and user behaviour will determine the degree of competition between competitive chatbots or answer engine services and more monopolistic search engine services.

This paper is structured as follows. Section 2 focuses on search engines. Section 2.1. reviews the existing empirical research literature on economies of scale in user data collection on search-engine efficiency. Section 2.2 discusses the impact of DMA search-engine data-sharing obligations on market structure and user welfare. Section 3 focuses on chatbots or answer engines. Section 3.1. explains the generative AI technology behind chatbots and the lack of user network effects in this technology. Section 3.2. examines substitution between search and chatbots in function of user welfare effects that depend on semantic complexity of queries. Section 3.3 examines potential impact channels on the advertising side of multi-side markets. Section 4 concludes with some competition policy recommendations.

2 Search engines

The search engine market is a near-monopoly, with Google holding a 90 percent market share and Microsoft Bing and Yahoo around 3 percent each[3]. A few very small players have negligible market shares. Google's dominant position in search has implications for the entire online services market. Google search captures early 80 percent of all desktop search and 90 percent of mobile search[4]. It has become an unavoidable gatekeeper to access many other online services. Google can leverage this position to extract rents from service providers, often in the form of side payments (paying for sponsored ads to stay on top of search), or to vertically integrate in some online services markets and

[2] This paper uses 'chatbots' and 'answer engines' as catch-all terms to represent generative AI models, such as generative pre-trained transformer (GPT), ChatGPT and related chatbot technologies, that apply Large Language Models to respond to queries.

[3] Source: Similarweb, March 2023, https://www.similarweb.com/engines/. Note that search-engine market share statistics vary. Statista for example, gives Google Search an 85 percent market share and Bing about 8 percent. See https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/.

[4] Source: https://www.businessdit.com/search-engines-usage-statistics/.

give preferential treatment to its own services. The Google Shopping case[5] is a typical example. Its strong market position in search also results in a monopolistic market position inside Google Search's 'walled' advertising garden. Moreover, the data that it collects from search powers a strong position in open advertising markets outside Google Search pages. In line with the economic logic of multi-sided markets, users get free access to search services, while advertisers pay to send targeted ads to users (Parker and Van Alstyne, 2005; Rochet and Tirole, 2006; Caillaud and Julien, 2004). Advertisers cannot multi-home. They have to be on the search platform if they want to target consumers there. This results in high advertising market entry costs, which finance the search engine business model. Consumers on the other hand benefit from substantial consumer surplus from free access to search services (Brynjolfsson and Eggers, 2019), though the economic impact of reduced consumer privacy is unknown.

In the logic of EU regulators, Google's dominance of search is due to direct and indirect number-driven and data-driven network effects in multi-sided markets: more end users in search attract more end users because more search data collection improves the quality of the search service. The cure for this dominance, as reflected in DMA Art 6(11), is to break these network effects. The instrument to achieve this is to redistribute the massive trove of user data that a dominant search engine collects, and share it with smaller search engines and new market entrants. This would neutralise the scale advantage of a dominant search engine, obtained through network effects, and result in more competition between search services.

In this section we first investigate empirically what the relevant research literature tells us about the workings of search engines and what the likely impact of search engine data sharing would be on all sides of the search platform market, including competitors, end users and advertisers. We then explore various options for the design of the data-sharing regime.

2.1 User data and efficiency in search engines

Adomavicius and Tuzhilin (2005) explained how search engines, or recommender systems in computer science parlance, work. The central problem of a search engine is to maximise consumer utility from a set of replies s_i, drawn from a total set of potential replies S, in response to a query from user i in a set of users C, each with individual characteristics or preferences c_i. A reply can be the URL of a webpage, or a picture, video, soundtrack or location on a map. S can range from relatively small, for instance the number of books in a store, to very large, such as the set of nearly 2 billion websites and 50 billion webpages on the internet[6]. Matching queries with replies is usually a statistical process; many replies may fit the query. How to find the best possible replies? Replies are usually not well-defined for the entire space C x S, only for a small subset of that space, ie subsets of users and items. The matching process then requires tools to extrapolate utility from the defined parts of that space where some answers are already known, to the undefined parts of that space. The authors identified two methods:

- Individual user-focused recommendations whereby the user is steered towards items in S that are similar to those preferred in the past. Past preferences may be a guide for future choices.

- Social or collaborative recommendations where the user receives recommendations based on what other people with similar preferences liked.

[5] Source: https://curia.europa.eu/juris/liste.jsf?num=T-612/17.
[6] Source: https://review42.com/resources/how-many-websites-are-there/.

In practice, most search engines will combine both methods and extract a subset of web URLs from S that combines indirect utility estimates from the user's own preferences c_i, as well as other users' preferences $\sum cj \neq i$. User-preference profiles can store information from the user's past searches and from other personal data sources outside the search engine.

Empty parts of the C x S space will gradually fill up as more users launch queries in the search engine and make their choices by clicking no web pages proposed by the search engine in the space S. The search engine returns a ranking of pages, and a snippet of their content, by relevance. Users click on top-ranked or lower-ranked items. Each click represents an appreciation of the contents of the page by a user. Items that are clicked more frequently by a user, or by other users, in response to a specific query, will move higher in the ranking.

In this way, the efficiency of a search engine, as measured by click-through rate (CTR) on higher-ranked items, is inherently subject to learning over time and across users in C and items in S. In other words, pooling of query and response data across users is a central feature of search. Translated into economic jargon, this refers to the importance of economies of scale and scope in data collection (Carballa *et al*, 2022). Economies of scale in search could be interpreted as the number of observed user clicks on a particular page in response to a particular query and ranking of that page. Economies of scope are reflected in the range of items S for which it has collected user information, or the range of different queries to which it can give a satisfactory response. Search efficiency, or the steepness of the search engine learning curve, relies on the collected pool of user data.

Several empirical papers have explored these learning dimensions and the characteristics of the learning curve, in particular with regard to economies of scale and scope.

McAfee *et al* (2015) explored differences in performance between large and small search engines, say Google Search and Bing: to what extent is performance related to scale? More than half of all queries are rare (see below for a definition of 'rare'). This creates a matching problem: which web page is most likely to contain an answer to a query? Statistically speaking, search engine web crawlers index nearly 2 billion websites with about 50 billion ($3x10^{10}$) webpages on the internet, with billions of word combinations. Each of these can contain a response to a query. Google collects about 8.5 billion queries and clicks per day[7], x 365 days = 3.1 trillion ($3.1x10^{12}$) observations of queries and clicks on 50 billion pages. That leaves a space with 9×10^{22} combinatorial possibilities and trillions of degrees of freedom to estimate the best page response to a query. How to narrow that set? Bing, Yahoo and Baidu collect much smaller numbers of user queries and clicks, proportional to their search market shares. Does this mean they are less efficient in matching? If economies of scale (number of users and uses) and scope (scope of data collected) in data aggregation matter, one would expect smaller search engines to be less efficient. Competition policymakers seem to think so, and therefore want to unwind these scale benefits. Do we have evidence to that effect? And what would be the consequences of reduced scale and scope?

McAfee *et al* (2015) employed several methods to answer that question:

- Rare query analysis: as more data on rare queries accumulates, the quality of search improves. The authors compared CTR on rare queries as the number of queries increases. They took benchmark data for rare queries in 1Q2014 (queries that occured fewer than 200 times in that period) and compared Google and Bing CTRs on these rare queries in 2-4Q2014, provided they remain relatively rare in that period (between 100 and 1000 queries in that period). They found

[7] See Maryam Mohsin, '10 Google search statistics you need to know in 2023', *Oberlo*, 13 January 2023, https://www.oberlo.com/blog/google-search-statistics.

that the CTR increases as the number of queries increases within the 100-1000 bracket. Scale matters.

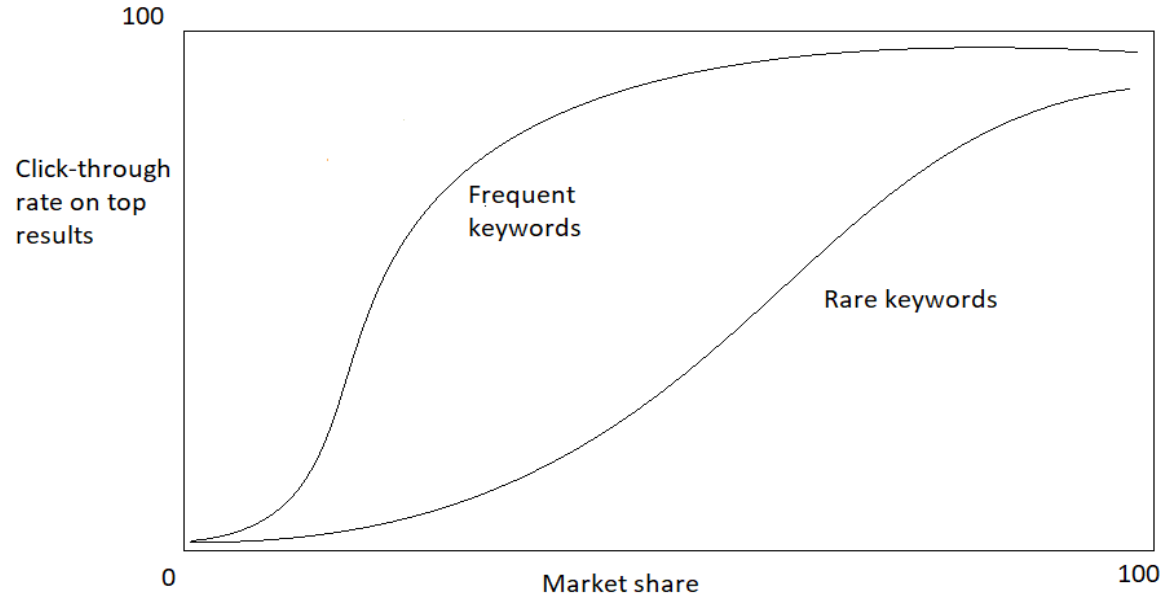
- Direct and indirect position analysis: related (indirect) queries can help improve the CTR on rare queries, though many rare queries have few related queries. As scale increases, the number of indirect or related queries increases and will improve the CTR. One hundred billion searches on Bing are mapped to 2.6 billion similar queries and 128 clusters of semantically related queries. Many queries have very limited indirect queries. Still, the authors showed again that scale matters: Bing gets better with more views, whether direct or indirect.

- The authors demonstrated that new web pages with better information content do not drive the CTR for rare queries. Improvements in CTR are driven by improvements in the quality of web page rankings as more direct view-count data moves relevant web pages further to the top of the search ranking.

He *et al* (2017) asked if search quality is determined mainly by algorithmic quality or by user clicks and feedback. In the first case, a well-funded new entrant could potentially produce results of superior quality compared to the entrenched market leader. In the second case, learning from historical user queries is critical to ranking quality. A technologically superior but data-starved algorithm would perform much worse than the incumbent's, unless the steep part of the learning curve with strongly positive returns occurs already at small user market shares, which any viable entrant could achieve easily. These two cases are very different in terms of the potential for innovation and competition dynamics.

He *et al* (2017) proved empirically that more common queries have higher CTRs and that the relationship is proportional to the square root of the log of historical clicks. This indicates an S-shaped learning curve in which increases are higher at lower data levels. They redid the McAfee *et al* (2015) analysis of rare query data and showed that as initially rare queries (<100 per month) grow moderately popular over time (>300 per month), the CTR increases by about 2-3 percent. Faster increasing returns to scale at low scale suggest that new entrants into the search market can rapidly achieve good quality search with small market shares. However, this is much more difficult for queries that remain rare over time – which may account for 25 to 50 percent of all daily queries in a search engine. They also redid the clustering and related query analysis of McAfee *et al* (2015). They showed that semantic clustering of similar but not identical queries can results in a ten-fold reduction in uniqueness of queries in the long tail, thereby providing an additional source of economies of scope in data aggregation. He *et al* (2017) concluded that both Google and Bing are running short of data in the long tail of queries, which accounts for about half of all queries. While both are well within the region where learning curves start to flatten out, with diminishing returns to additional data, the slope is still positive. They would both benefit from more data. The authors suggested that major algorithmic improvements could still overrun the data advantage. Figure 1 summarises these findings: steep search engine learning curves, followed by diminishing returns, for frequent queries that can occur at low search market shares; and flatter learning curves for rare queries that require a larger market share to collect sufficient data to keep the learning curve rising.

Figure 1: Search engine learning curves by frequency of keyword queries



Source: Bruegel. Note: CTR = 100 includes cases where users are satisfied with the replies on the first search results page and do not click on any page link.

While He *et al* (2017) used direct and indirect query results generated by users, Schaefer and Sapi (2019, 2022) added a new source of information, user history. If a user uses the same search engine over a longer period of time, the engine collects more data on user preferences and keywords, enabling it to respond better to that user's queries. Using data from the Yahoo search engine, they first confirmed the positive relationship between search quality (CTR on the top-ranked page) and the total number of searches for a keyword. Second, they found that the increase in search quality from additional searches was more pronounced as the average user history increased. This suggests that more user-specific information, and more information from users with similar profiles, makes learning in the across-user dimension more efficient. Applying the jargon of He *et al* (2017), we could say that direct and indirect user query data can be clustered to generate more information on user profiles. More generally, Schaefer and Sapi (2022) concluded that additional observable user characteristics, possibly from other sources than search queries and clicks, can help to improve the performance of a search engine. This could be interpreted as an additional source of economies of scope in search engines: adding more explanatory variables from different sources increases the accuracy of prediction of the desired user result. They found an S-shaped curve with increasing and diminishing returns to data scale.

However, like He *et al* (2017), Schaefer and Sapi (2022) were reluctant to extrapolate their results to real-world search markets. They noted that their sample search data was difficult to relate to real search traffic and therefore hindered an exact identification of scale effect sizes. In other words, the location of the switching points to increasing and decreasing returns are hard to pinpoint and depend on the accumulated history of query-specific and user-specific data.

Schaefer and Sapi (2022) explored the implications of their findings for competition policy and digital regulation. Since personalised data increases the efficiency from learning across different users, merging databases across services with overlapping users may lead to significant efficiency gains (Eisenmann *et al*, 2011). The importance of personal data also suggests that sharing non-personal

search, query and keyword data, an obligation imposed by the EU Digital Markets Act on very large 'gatekeeper' search engines, may not be sufficient to restore a level playing field among competing search engines, because that data cannot be connected to individual users. Anonymisation will reduce the efficiency of search and will be less effective in terms of fostering competition between search engines.

Building on an earlier search engine model by Argenton and Prüfer (2012), Prüfer and Schotmuller (2022) studied data-driven search engine markets where the marginal cost of service quality is decreasing in the amount of user data, generated as a byproduct of using a service. This implies that higher sales volumes or market shares today reduce the cost of satisfying users' preferences tomorrow. The authors constructed a theoretical model of dynamic R&D competition, in which duopolistic competitors repeatedly and sequentially chose their rates of innovation. The model incorporates data-driven indirect network effects: more users on the demand side, and the data that they generate, reduce the marginal cost of data-based innovation on the supply side. They showed that this type of market will eventually tip and one firm will dominate. The weaker firm will never acquire more than a negligible market share in the future. Beyond this tipping point, neither the dominant nor the weaker firm has incentives to invest further in innovation. It deters market entry even for innovative firms. Innovation stalls, which is bad for consumers. According to the authors, only mandatory data sharing between the dominant and small search engine could overcome this problem. Both competitors would face the same data cost function and could compete effectively.

Klein *et al* (2022) presented experimental empirical evidence in support of data-driven network effects and dominance in search engine services markets. Rather than using CTR, the authors let humans assess the quality of responses to keyword queries. The assessment found that larger search engines produce better quality results. This can be due to better algorithms or more data. To answer this question, the authors split keyword queries made to the small search engine into five groups, by popularity of each query. The most popular keywords accounted for only 0.2 percent of all queries but 11 percent of traffic. The least popular keywords accounted for 75 percent of all queries and 56 percent of all traffic on the search site. To test the impact of data availability on the quality of search engine responses, while keeping algorithms unchanged, the authors artificially reduced the amount of user data input into the search engine to produce responses to specific queries. Data inputs were split into ten buckets of 10 percent of all data, and the number of buckets fed to the search engine was gradually increased to 100. They complemented these small search engine results with non-personalised search results from Google and Bing for the same queries. The findings point towards first increasing, then decreasing, returns to scale in data inputs. Data-scale effects are weaker for popular queries, more pronounced for rare queries. Differences in data inputs do make a difference, for unchanged algorithms. The authors concluded from this experiment that mandatory data sharing, as foreseen under the EU DMA, would indeed allow new entrants and small search engines to compete more effectively with incumbents that dominate the market.

Kramer (2023) examined practical obstacles to search-engine data sharing: privacy risks and costs. As explained above, Schaefer and Sapi (2022) showed that personal data is essential for efficient search. Krämer (2023) explored the tension between search-engine data sharing and privacy protection. Sharing detailed query, response and clicks datasets entails privacy risks. Even if anonymised, there are risks of de-anonymisation. Several techniques to reduce that risk, for example aggregated data, k-anonymisation or adding noise to data, make the data less useful for competing with incumbent search engines. Apart from technical means, there may be institutional solutions to improve the efficiency-privacy trade-off: storing data in intermediary trusts (Prüfer and Graef, 2020), creating privacy sandboxes with in-situ access (Martens *et al*, 2021), etc.

Another obstacle is the huge amount of search data to be shared. Google generates about 3 trillion searches per year. This data is very costly to store and make available for third-party processing. Detailed data sharing may also enable reverse-engineering of gatekeepers' algorithms. Sharing would be more feasible if it would be limited to subsets and to recent search data only. However, subsets reduce data density in the long-tail of rare queries, which is precisely where smaller search engines lack sufficient data (Klein *et al*, 2022; He *et al*, 2017). Subsets may therefore be sampled with a bias towards long-tail rare keyword searches. Another solution is to limit shared data to top ranked results, or to clicked results only.

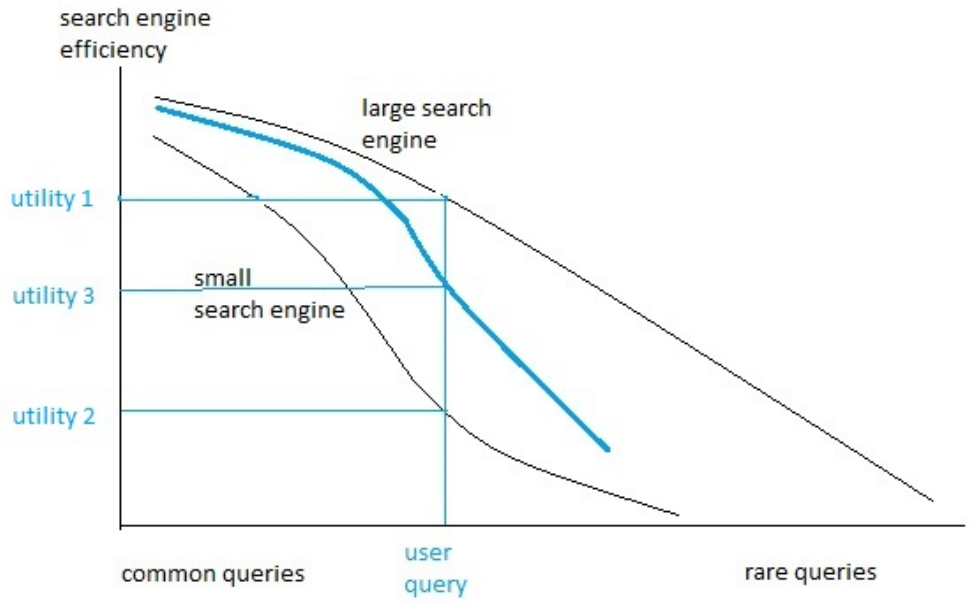
2.2 Discussion

All empirical papers conclude that search-engine efficiency is driven by network effects and economies of scale and scope in user data aggregation. However, an unresolved problem is to estimate the points at which increasing and decreasing returns to scale kick in at market level, not only at sample level in the studies. He *et al* (2017) pointed out that a smaller search engine like Bing runs into problems predicting replies to rare keyword queries. Google's sheer scale makes it more efficient for rare keywords, though there may still be positive returns to scale even at Google's market share of 90 percent. We do not know how much the efficiency of Google Search for rare keywords would decline if its overall search market share decreased from 90 percent to, say, 50 or even 20 percent. To the best of our knowledge, such market share 'stress' tests have not been tried for Google Search.

This is a crucial issue though. The EU Digital Market Act forces Google to share its search data with smaller search engines. Google Search will lose market share and collect less user data. This will result in lower search efficiency for Google. Smaller search engines will pick up Google data and can combine them with their own data. But if both datasets have lower quality compared to Google's original search data before its market share declined, than the overall combined efficiency of all search engines will decrease.

This is reflected in Figure 2 below. Consider a market with two search engines, a small and a large search engine. The latter is more efficient because it collects more user data. A user who launches a random query will achieve maximum utility 1 on the large search engine and a lower utility 2 on the small search engine. When the DMA forces the large search engine to share its data with small search engine, the efficiency of the first will be reduced because it loses market share while the efficiency of the second will increase. When efficiency converges in the blue line, somewhere in between the two original curves, the same user with the same random query will achieve utility level 3 on both search engines. The reduction in utility will be minimal for common queries but increase for rare queries. Users cannot reach utility level 1 anymore by staying with the large search engine because that loses users and thus data, especially for rare queries. User welfare thus diminishes with asymmetric data sharing. We do not know yet from the available studies at what critical level of market shares this happens, or for which frequency of keyword queries and how much search-engine efficiency decreases. We know however that it is very likely to happen at some degree of market fragmentation.

Figure 2: Search engine data sharing



Source: Bruegel.

This illustrates the tension between traditional competition policy and data-driven network effects in large platforms. Cabral *et al* (2021) already pointed out that promoting competition and market fragmentation, as the DMA seeks to achieve in search engines, may come at the expense of reduced network effects and a consequent reduction in user welfare. Network effects may have a negative impact on competition. But network effects also represent positive externalities that increase the welfare of users. Trying to find an appropriate balance between the positive welfare effects of increased competition and negative effects of reduced network externalities is a difficult task for policymakers.

Another way to approach this is to start from the idea that search data aggregation has social value for all search-engine users. In the current market situation, one search engine, Google Search, does that aggregation because it covers most of the market. Bringing more competition into the market weakens Google's ability to collect relevant market data. Alternatively, we can try to restore the social value of aggregated search data, even in the presence of many competing search engines with small market shares. This could be achieved by pooling data from all search engines in a single data pool managed by a third-party (Prufer and Graef, 2020), or by requiring mutual data sharing between all search engines, irrespective of size or market share. Even small search engines with small market shares can hold a few valuable datapoints for rare queries. If pooled together across all search engines, this may constitute a sufficiently large pool in which no observation on rare queries gets lost. In terms of Figure 2, the blue line would then represent the learning curve at full 100 percent market share. Unfortunately, third-party data pooling and symmetric data sharing are not foreseen in the DMA obligations.

Krämer (2023) pointed out the cost of sharing a massive volume of search data. Tens of billions of queries on Google Search every day generate a huge volume of data to be duplicated and transferred between search-engine providers. He suggests that in order to cope with the huge volume of search engine data, the scope of data sharing could be limited to (a) rare keyword queries that matter most for

small search engines, (b) to responses and clicks for top-ranked or first-ranked pages only, and (c) reduce it to clicked pages only.

Krämer (2023) also alerts us to the risks of sharing personal data, unless by consent. Sharing anonymised data is always subject to the risk of de-anonymisation. In contrast, Schaefer and Sapi (2022) argued that sharing of non-personal data only, as regulated by the EU Digital Markets Act, is not enough because the personal-data dimension matters for search efficiency. Since user search histories and other user profile data can contribute to the efficiency of search, search engines could create the possibility for consumers to port their search histories and personal profile data to other search engines – a possibility that is foreseen in the EU General Data Protection Regulation (2018), but not in terms of real-time transfers.

3 From search to chatbots

In this section, we switch from search engines to answer engines or chatbots, based on LLM technology. An answer engine does not match users with webpages that may contain an answer to a query. Rather, it provides a full semantic answer to the query. In this section, we investigate the impact of this technological change on the quality of answers to queries and on market structure in search services.

3.1 The technology behind generative AI and Large Language Models

Major breakthroughs have occurred in AI neural network and deep learning algorithms in the last couple of years, notably with the 'transformer' models produced by Google researchers (Vaswani *et al*, 2017). This has enabled a vast expansion in the scale and generalisation of AI models and led to so-called *"foundation models"* (Bommasani *et al*, 2020). Previous generations of deep learning models required datasets annotated by humans for learning a specific task. The cost of annotation imposed limits on the amount of data available. Narrow training datasets made it difficult to apply the model outside the training domain. Foundation models are pre-trained on much larger but unannotated datasets. LLMs used in answer engines, are a subset of foundational models that focus on natural language processing. Foundation models are very versatile and can be applied to a wide range of tasks, even outside the domain on which they were trained. They are general multi-purpose models that can be applied to language, image recognition and audio-visual tasks, molecular structures and computer programming. This trend towards *homogenisation* or application of a single model to a variety of tasks also carries risks that biases learned during training are transferred to applications in a wide variety of domains.

Foundation models like LLMs require very large scale computing resources. The earliest GPT models took about 0.1 percent of the resources required for the current GPT4 generation (OpenAI, 2023). Sevilla *et al* (2022) documented how training compute capacity has grown exponentially. Before the arrival of deep learning AI, compute doubled every 20 months. Since the arrival of deep learning in 2010, it doubles every six months. LLMs training costs can run into millions of dollars (Bowman, 2023)[8]. Pre-training of LLMs is done on very large sets of text data, usually taken from books, Wikipedia and webpages[9]. They can also be pre-trained on audio or image data. The earliest models used datasets with billions of words as training inputs; current models use trillions.

[8] The high cost of training LLMs explains why most LLM research and development is now done at large firms, rather than academic institutions, that have the required budgets and computing infrastructure. A consequence of this 'privatisation' of AI research is that technical details of new models and training conditions are often no longer published.
[9] See for example Kindra Cooper, 'OpenAI GPT-3: Everything You Need to Know', *Springboard*, 1 November 2021, https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/.

The very large scale of the models led to unexpected *emerging* properties, the ability to learn in-context with a natural language *prompt* that describes the task with an example, a reasoning how to address the task or explaining the similarity with other tasks. This is called 'few-shot' or 'zero-shot' learning and it works with similarity functions that compare the characteristics of new data with previous collected data[10] (Kundu, 2022). It enables a pre-trained model to generalise over new categories of data that were not available during training (Zhao, 2022). Prompted few-shot and zero-shot learning avoid the high cost of re-training an LLM. This has been used to 'correct' LLM model responses through reinforcement learning from human feedback (RLHF) (Bai *et al*, 2022). This can also be used to correct for 'hallucinations', when models produce erroneous output that is not based on training data, or to align output with human preferences for harmless and helpful output. LLM models may interact with selected human agents who give feedback to the model, either with examples of preferred reasoning or with an evaluation of the output. LLM developers decide which feedback to incorporate into the model. The user feedback curation role of developers is crucial to prevent users from tricking an LLM into false or harmful pathways. Note that user feedback is not a statistical correction mechanism, as it is in search engines; it does not rely on a statistically relevant sample of user feedback data.

Prompts enable LLMs to learn how to respond to rare or highly specialised queries, even when they had only limited or no exposure to these queries during training. 'Prompt engineering' has become a new discipline that enables users to feed an LLM with a set of appropriately formulated prompts to guide it to the correct answers and avoid hallucinations[11]. 'Grounding' techniques are used to let an LLM take into account information contained in proprietary datasets[12]. That avoids very costly re-training of an LLM to incorporate new data, an option that is usually not feasible for smaller firms or individual researchers. 'Plugins' can add specific sets of prompts and application-specific datasets to an existing LLM answer machine to make it more suitable for specific applications[13]. The versatility of LLMs to applications in a wide range of areas, often unconnected to the data pool on which they were trained, makes them very powerful for many real-world applications. In fact, LLMs as natural language processing models are often seen as a subset of a wider category of foundation models that do not necessarily use natural language (CMA, 2023). The ability of LLMs to respond to single-shot prompts is due to the rapidly expanding scale of training data and model size that have led to unexpected emerging new abilities that were not observed at smaller scale. For example, Mueninghoff *et al* (2022) found that LLMs are capable of zero-shot generalisation to tasks in languages they have never intentionally seen. They conjecture that the models are meta-learning higher-level capabilities that are both task- and language-agnostic. Bowman (2023) remarked that: "*These results are in tension with the common intuition that LLMs are nothing but statistical next-word predictors, and therefore cannot learn or reason about anything but text. While the premise of this intuition is technically correct in some cases, it can paint a misleading picture of the often-rich representations of the world that LLMs develop as they are trained*".

While search engines still rely on statistically representative sets of user data to produce a meaningful output, especially for rare queries, very large-scale answer engine LLMs are no longer statistical matching machines, at least not at the semantic level of queries and responses. That explains why user-generated data-driven network effects do not play a role in answer engines. LLMs are up-and-

[10] Rohit Kundu, 'Everything you need to know about Few-Shot Learning', *Paperspace Blog*, undated, https://blog.paperspace.com/few-shot-learning/.
[11] See https://github.com/dair-ai/Prompt-Engineering-Guide.
[12] Mick Vleeshouwer, 'How to create a private ChatGPT with your own data', *Medium*, 27 March 2023, https://medium.com/@imicknl/how-to-create-a-private-chatgpt-with-your-own-data-15754e6378a1.
[13] Mick Vleeshouwer, 'How do ChatGPT plugins (and similar LLM concepts) work?' *Medium*, 3 May 2023 https://medium.com/@imicknl/how-do-chatgpt-plugins-and-similar-llm-concepts-work-2c83a4aeedd4.

running from the first user onwards. Answer engines have contextual memory that keeps track of user responses and prompt instructions during a conversation. This memory is not carried over to the next conversation. LLM developers may however decide to incorporate user feedback to correct for hallucinations and erroneous answers. The LLM cannot learn directly from interaction with users. That would expose the model to learning malicious or incorrect content from biased users. This user feedback mechanism is very different from the reliance of search engines on user clicks on ranked pages.

The mechanisms that drive these emerging properties are still somewhat unclear. They make it very difficult to 'explain' in a transparent way what is happening inside an LLM. Bowman (2023) commented that "*As of early 2023, there is no technique that would allow us to lay out in any satisfactory way what kinds of knowledge, reasoning, or goals a model is using when it produces some output*".

3.2 The impact of answer engines on market structure in search services

From a competition-policy perspective, a crucial characteristic of chatbots or answer engines is the absence of direct data- and number-driven network effects. That reduces the risk of user lock-in and a monopolistic market structure. Economies of scale and scope in data aggregation occur at the pre-training stage of a LLM, not during interaction with users.

This facilitates market entry by new players. Small search engines, including DuckDuckGo, Yahoo, You.com and Neeva, that already rely to some extent on access to Microsoft Bing's index of webpages, have now introduced chatbot functions in their search engines. This enables them to bypass network effects to rapidly improve the efficiency and coverage of their search engines, beyond the efficiency levels that they can learn from their limited sets of users, especially for rare queries. The arrival of generative AI LLM models has stimulated frenetic investment and innovation activity by incumbents. The dominant incumbent search engine, Google Search, is trying to catch up rapidly in technology with the main challenger, Microsoft Bing, supported by OpenAI's ChatGPT LLM. Google launched the competing BERT and BART LLM answer engines, initially to mixed reviews. It now tries to leverage its vast trove of consumer data to get a stronghold in the answer-engine market. Rapid evolution in LLM prompt engineering, specific data grounding and plugin technologies are likely to result in the emergence of many variations on LLM models that make more use of users' own and other users' data. All of these offshoots may become competitors of search engines in specific niche markets.

We discuss the impact on consumers in the next section. In this section we focus on market entry and the supply side of search and answer engines. Market entry is conditional on access to key inputs for training of LLMs. Training LLMs requires two inputs: (a) training data inputs from a very large body of text, usually collected from an index of web pages by web crawlers, books, documents, etc, and (b) training infrastructural inputs including hyperscale cloud and computational infrastructure to run the training sessions:

(a) Data: Web page indexes are available as a commercial product from Google Search and Microsoft Bing only. Microsoft apparently opposes the use of its indexes by competing chatbot service providers[14]. It claims that this use goes beyond the terms and conditions of the agreement to use Bing web-crawler data. Webpages need to be scanned to collect text data input to train the LLM model. In addition to webpages, LLM models can use text documents collected elsewhere. LLM models can also be trained on images (pixel patterns) and sounds (soundwave patterns) to generate new images,

[14] Emma Roth, 'Microsoft reportedly orders AI chatbot rivals to stop using Bing's search data', *The Verge*, 25 March 2023, https://www.theverge.com/2023/3/25/23656336/microsoft-chatbot-rivals-stop-using-bing-search-index.

music and speech. Text and audio-visual media may be subject to copyright protection. There is a hot academic debate on the use of copyright-protected material for LLM training purposes (Appel, 2023; Gilotte, 2020). Several relevant court cases are on-going in the United States and the EU. National copyright law and the EU Copyright Directive are not clear on this issue. If rightsholders need to be paid for the use of their content, managing the rights and remuneration of a very large number of rightsholders may run into very high transaction costs that may exceed the value of the remuneration. Payments may disincentivise the use of LLM models.

(b) Infrastructure: Hyperscale cloud and computational infrastructure is available only from a few very big cloud computing services suppliers. This explains why companies that already own hyperscale cloud infrastructure, such as Amazon and Baidu, are in a position to rapidly start up chatbots. The combination of inherently high fixed training costs for LLM models and oligopolistic inputs markets may constitute an obstacle to market entry for start-up chatbots or answer engines. Research on building LLMs and wider sets of foundational models is located almost exclusively in a few big tech companies. Academic institutions are not in a position to finance the resources required for training these models (Bommasani *et al*, 2022, p 10). Centralisation of LLMs in big tech firms also results in the privatisation of AI research, pulling it out of the public domain. Many trained models are not released in the public domain[15]. Properties like in-context learning only emerge in very large scale models that are beyond the resource constraints of academic institutes. Academics cannot studies these properties to get better insights into the mechanisms and reliability of these properties.

While the absence of user-driven network effects facilitates market entry for smaller start-ups and competition in answer engine services, very extensive data and computing resources requirements *de facto* limit competition to a relatively small set of big tech companies and a few start-ups with deep financial pockets to acquire the necessary inputs. Inputs markets are relatively oligopolistic and will probably results in a few players having sufficient resources to invest in the training of foundation LLM models (CMA, 2023). They may collaborate with a wide variety of specialised off-shoots that adapt these models to specific applications. Still, even with a few big players offering foundation models, surrounded by many smaller specialised players, the market structure for LLM answer engines is likely to be very different from the current search market structure in which a single dominant player occupies more than 90 percent of the market.

3.3 The demand side

Because answer engines predict sequences of words rather than URLs, they can generate semantically more complex replies compared to search engines. Search-engine users need to construct the semantics of a reply to their query from scanning and reading through the list of web pages that they receive. That imposes a cognitive cost on users, the transaction cost of finding the desired reply. Answer engines take over a substantial part of the cognitive processing work that users would do when working with a search engine. That reduces cognitive processing costs for users. The corollary of this reduction in user search costs is an increase in *ex-post* uncertainty for users: is the reply correct, what sources have been used to produce this answer, how credible are these sources and the extracted answer, have some relevant sources been overlooked, etc? There are already many stories about chatbots producing nonsensical or wrong replies. The pre-processed reply from a chatbot reduces transparency compared to a user-produced reply based on webpage output provided by the search engine.

[15] Nitasha Tiku and Gerrit De Vynck, 'Google shared AI knowledge with the world — until ChatGPT caught up', *The Washington Post*, 4 May 2023 https://www.washingtonpost.com/technology/2023/05/04/google-ai-stop-sharing-research/.

The cognition-augmenting impact of chatbots has been observed in a real-life work setting. Brynjolfsson *et al* (2023) studied the impact of chatbot assistants for call-centre workers. They found an average productivity increase of 14 percent, mainly for lower-skilled and less-experienced workers who still need to accumulate a sufficient stock of tacit knowledge to perform well on their tasks.

Incentives for users to substitute search engines for answer engines will depend on the semantic complexity of a query. For example, if a user is looking for the address of a restaurant or the opening hours of a shop, both search and answer engines will deliver a semantically simple and easily verifiable reply. In this case, there is no significant difference in user transaction costs between the two. If a user seeks for example an analysis of a political system in a country, the reply will be semantically more complex. Answer engines will do a lot of the cognitive processing but may black-box sources and credibility. Search engines on the other hand may at best generate some web pages with information on the political system that the user will have to digest to make her own assessment of the credibility and possible bias in the information sources. The degree of substitution between search and answer engines thus depends on the cognitive complexity of the question, users' opportunity cost of cognitive processing time and their perceived risk aversion.

Incumbent search engines Google Search and Microsoft Bing respond to this substitution effect by providing search and answer engine responses to a query in parallel on the same page. For example in Microsoft's revamped Bing search engine, called Prometheus[16], the replies generated by Bing search and by the chatbot are combined by means of an 'orchestrator' function into a single response page. Users can then select or combine both search and answer engine replies. While reading a longer reply from the answer engine, they may still click on links to relevant information source webpages and verify the credibility of the reply. Search and answer engines may thus become complements rather than substitutes. New market entrants who do not have a scaled-up and well-performing search engine can only offer answer engines replies.

Unlike Microsoft Bing, smaller start-up answer engines like NeevaAI do not have user data on query-relevant webpages to combine with their answers. They use the accumulated stack of independently crawled webpages – that they need to feed the chatbot – to estimate webpage relevance on the basis of incoming links in pages, and displays the most relevant pages together with chatbot answers[17]. It is no guarantee that users find these links relevant, but it may well be a reasonable proxy.

Cost considerations may also play a role in the competition between search and answer engines. Answer engines are more expensive to run, especially for semantically complex queries that require longer answers. For simple queries, service providers may prefer to rely more on search engine output. Some answer-engine providers, such as Microsoft Bing, limit the amount of text output in order to reduce processing costs. In future, subscription-based services may shift these costs to users. It is not clear to what extent the zero-priced user side in search engines, financed by positive-priced entry on the advertising side of the platform, can be maintained for answer engines (see next section).

Partial substitution of consumer demand between search and answer engines raises the question of whether the current monopolistic search market structure will weaken or even disappear with the emergence of a more competitive market for answer engines, because these are not subject to user network effects – though oligopolistic constraints in input markets may still limit competition in answer-engine services. If answer engines can effectively outcompete search engines, the DMA's search-engine data sharing obligations may no longer be necessary to restore a competitive market.

[16] See https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing.
[17] See https://neeva.com/blog/introducing-neevaai.

Technological innovation may be able to achieve that goal, without regulatory intervention in data markets.

3.4 The advertising side

In multi-sided platform markets, one can never evaluate the changes on one side without looking at the impact on the other side. That brings us to the other side of search platforms: online advertising. Google Search runs on a business model financed by advertising. Users do not pay for access to search services; advertisers do.

The introduction of chatbots or answer engines can trigger at least two effects on advertising:

First, there may be an impact on user engagement with the platform, including with ads. In Google Search, about a quarter of all searches stops on the first search results page, without click-throughs to other web pages[18]. With chatbots, the share of zero-searches is likely to increase because users find a semantically more elaborate and satisfactory answer to their question on the first page. Referral traffic from answer engines to other webpages, and advertising opportunities on these pages, will decrease. Users may however stay longer on the answer page. That may open more opportunities for revolving advertising on that page. In both cases, the relative value of advertising in the 'walled' garden of answer-engine pages is likely to increase compared to advertising on related pages.

Second, LLM models may affect the efficiency of targeted advertising. Meta and Google hold vast amounts of consumer data that can be used in combination with LLM models to increase the efficiency of targeted advertising. Advertising LLMs can be grounded with user profile data to build more targeted ads. Users could receive their own unique ads that take into account their personal preferences. Meta is using LLM models to design and implement advertising campaigns[19]. The corollary of this evolution is that advertisers will have to delegate a substantial part of the design of their advertising campaigns to proprietary LLMs. That will increase the degree of vertical integration of ad publishers, not only in walled online gardens but also in open web advertising.

Uncertainty about the future of advertising revenue, combined with the high cost of setting up and running answer engines, casts doubts on the financial sustainability of the current business model of search engines, which rely on advertising to cover the cost of free search. A shift towards a subscription-based business models for answer engines, especially domain-specific or firm-specific answer engines, may be required. Microsoft is well-placed in this respect because it could add the Bing answer engine to its subscription-based package of its Teams and Office productivity software.

4 Conclusions and policy recommendations

We are now in a position to answer the two-world question raised by He *et al* (2017): do we live in a world in which search efficiency is entirely dependent on user data, irrespective of algorithmic improvements, or in a world in which search efficiency is independent of user data and driven entirely by algorithmic improvements? Traditional search engine technology clearly belongs to the first world, while LLM answer engines or chatbots belong to the second world. At the moment, we live in a world in which both technologies exist in parallel, though with different market structures. Search engines are

[18] Source: https://www.semrush.com/blog/avoiding-zero-click-searches/. This suggests that the CTR is not an ideal measure of search engine efficiency. A zero CTR on top-ranked results may also indicate high efficiency, especially for simple searches where users are looking for a small piece of information, like a name, email or location, that can be displayed in a short text snippet extracted from a webpage.
[19] See https://ai.facebook.com/blog/large-language-model-llama-meta-ai/ and https://yourstory.com/2023/04/meta-ai-powered-ads-game-changer-businesses.

subject to network effects, driven by the number of users and the amount of user data collected, resulting in the emergence of a dominant search engine – Google Search. Search engines need to climb a social learning curve: users benefit from the same and related keyword queries and clicks by other users. The climb is steeper for common than for rare keywords. Scale matters for search engines, not for answer engines that do not depend on user data. Market entry is easier because they can climb to the top of the learning curve during their training phase, before being released to users. Training costs are high however and require access to monopolistic inputs. Search engines can learn from individual profiles and search histories; answer engines are not focused on personal data.

The EU DMA obligations for large gatekeeper platforms prescribe data sharing between search engines as a remedy for a monopolistic search-market structure. This paper showed the risks of an asymmetric data sharing strategy, from large to small search engines only. Such a strategy fragments not only the market for search services but also the social value of search data, with the risk of reducing the efficiency and user welfare from search. Admittedly, the precise cut-off points at which increasing and decreasing returns to scale in the search-engine learning curve set in are still not well understood, especially for rare keywords. An appropriate data-governance regime with symmetric data sharing between all search engines, irrespective of size, could preserve economies of scale and scope in data aggregation across all search services providers.

Chatbots or answer engines may offer another route towards competition in search services. They are not subject to scale and network effects. Market entry is easier, provided high fixed costs of training LLM models and input supply bottlenecks can be overcome. Answer-engine outputs are partial substitutes for search-engine outputs. Answer engines reduce users' cognitive processing costs for complex queries, less so for semantically simple queries. A competitive market for answer engines may be in a position to outcompete monopolistic search engines, provided there is sufficient substitution between the two, with users switching to answer engines and providers being able to reduce the cost of running LLMs for semantically simple queries.

Based on the analysis in this paper, a set of policy recommendations for EU DMA policymakers is proposed below. Since DMA obligations for gatekeeper search engines apply only in the EU market, it will be interesting to observe the differential impact of (the absence of) mandatory data sharing in the US and EU search markets, to isolate the impact of pure technological competition between search and answer engines in the US from the added factor of data sharing in the EU market. There may however be cross-over effects of search data sharing from the EU to the US market.

Policy recommendations:

Search engines:

1. Start by implementing the DMA search engine data sharing regimes as foreseen in DMA Art 6(11), with asymmetric data sharing from the largest gatekeeper search engines to smaller search engines.

2. Define the scope of search engine data to be shared. If the volume of query, ranking and click data is high and costly to share, it may be reduced to relatively rare queries, top-ranked websites and click data only, without significant loss of efficiency.

3. Enable users to port their personal profile data to other search engines in real time.

4. Monitor the search-engine market and search-engine efficiency (as measured by the CTR on top-ranked search pages) for all search engines and at full market scale (not samples). If market entry

and fragmentation results in a detectable reduction in search-engine efficiency for the largest search engines, especially for rare keywords, the data sharing regime should be modified.

5. In order to increase search-engine competition without diminishing economies of scale and scope in aggregated user data, policymakers could require symmetric or mutual data sharing between all search engines, irrespective of size and/or gatekeeper status. This policy option goes beyond the DMA.

Chatbots and answer engines:

6. Facilitate access to webpage indexes collected by search incumbents. Regulators could extend the FRAND (fair, reasonable and non-discriminatory) pricing provisions for search engine data in the DMA to webpage indexes. Pricing should be sufficiently low to facilitate market entry for chatbots and sufficiently high to incentivise webpage indexing.

7. Facilitate access to audio-visual media content protected by IPR for the purpose of training LLM models. Regulators should re-assess the economic balance between static investment incentives and dynamic welfare benefits from IPR protection in the case of LLM training models, for society as a whole, not for private interests only.

8. Monitor the degree of substitution and complementarity of search and answer engines, especially when the two services are offered jointly on the same page. Substitution is likely to stimulate competition from pure answer engines. Complementarity may strengthen the market position of incumbent search engines and preserve network-effect driven market dominance in search.

References

Adomavicius, G. and A. Tuzhilin (2005) 'Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions', *IEEE Transactions on Knowledge and Data Engineering* 17(6)

Appel, G., J. Neelbauer and D. Schweidel (2023) 'Generative AI Has an Intellectual Property Problem', *Harvard Business Review*, April

Argenton, C. and J. Prüfer (2012) 'Search Engine Competition with Network Externalities,' *Journal of Competition Law and Economics* 8(1): 73–105

Bommasani, R., D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx ... K. Zhou (2022) 'On the Opportunities and Risks of Foundation Models', Center for Research on Foundation Models, Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, available at https://arxiv.org/abs/2108.07258

Bowman, S. (2023) 'Eight Things to Know about Large Language Models', mimeo, available at https://arxiv.org/abs/2304.00612

Brynjolfsson, E., A. Collis and F. Eggers (2019) 'Using massive online choice experiments to measure changes in well-being', *PNAS* 116(15): 7250-7255

Brynjolfsson, E., D. Li and L. Raymond (2023) 'Generative AI at work', *NBER Working Paper* 31161, National Bureau of Economic Research

Cabral, L., J. Haucap, G. Parker, G. Petropoulos, T. Valletti and M. Van Alstyne (2021) *The EU Digital Markets Act, A Report from a Panel of Economic Experts*, European Commission Joint Research Centre, Luxembourg: Publications Office of the European Union

Caillaud, B. and B. Jullien (2003) 'Chicken & egg: Competition among intermediation service providers', *The RAND Journal of Economics* 34(2): 309-328

Carballa, B., N. Duch-Brown, S. Hacuk, P. Kumar, B. Martens, J. Mulder and P. Prüfer (2023) 'Economies of scope in data aggregation: the case of health data', *TILEC Discussion Paper* 2023-3, Tilburg University

CMA (2023) *AI Foundation Models: Initial review*, UK Competition and Markets Authority

Deutscher, E. (2021) 'Google Shopping and the Quest for a Legal Test for Self-preferencing Under Article 102 TFEU', *European Papers* 6(3)

Digital Competition Expert Panel (2019) *Unlocking digital competition*, HM Treasury, available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf

Eisenmann, T., G. Parker and M. Van Alstyne (2011) 'Platform envelopment', *Strategic Management Journal* 32(12)

Gillotte, J.L. (2020) 'Copyright infringement in ai-generated artworks', *UC Davis Law Review* 53(5): 2655-2692

He, D., A. Kannan, T.Y. Liu, R.P. McAfee, T. Qin and J.M. Rao (2017) 'Scale Effects in Web Search', in N.R. Devanur and P. Lu (eds) *Web and Internet Economics*, Lecture Notes in Computer Science volume 10660, Springer

Kaplan, J., S. McCandlish, T. Henighan, T. Brown, B. Chess, R. Child ... D. Amodei (2020) 'Scaling Laws for Neural Language Models', mimeo, available at https://arxiv.org/abs/2001.08361

Klein, T., M. Kurmangaliyeva, J. Prüfer and P. Prüfer (2022) 'How important are user-generated data for search result quality? Experimental evidence', *TILEC Discussion Paper 2022-016*, Tilberg University

Krämer, J. (2023) 'Data access for search engines', in A. De Streel (ed) *Effective and proportionate implementation of the DMA*, Centre on Regulation in Europe

Martens, B., G. Parker, G. Petropoulos and M. Van Alstyne (2021) 'Towards Efficient Information Sharing in Network Markets', *TILEC Discussion Paper* 2021-014, Tilberg University

McAfee, P., J. Rao, A. Kannan, D. He, T. Qin, and T.-Y. Liu (2015) 'Measuring Scale Economies in Search', presentation to Lear Conference 2015, available at https://www.learconference2015.com/wp-content/uploads/2014/11/McAfee-slides.pdf

Muennighoff, N., T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao ... C. Raffel (2022) 'Cross-lingual Generalization through Multitask Finetuning', mimeo, available at https://arxiv.org/abs/2211.01786

Ouyang L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin ... R. Lowe (2022) 'Training language models to follow instructions with human feedback', mimeo, available at https://arxiv.org/abs/2203.02155

Parker, G. and M. Van Alstyne (2005) 'Two-Sided Network Effects: A Theory of Information Product Design', *Management Science* 51(10): 1494-1504

Prüfer J. and C. Schottmüller (2022) 'Competing with big data', *The Journal of Industrial Economics* LXIX(4)

Rochet, J.-C. and J. Tirole (2006) 'Two-sided markets: a progress report', *The RAND Journal of Economics* 37(3): 645-667

Schaefer, M. and G. Sapi (2019) 'Data Network Effects: The Example of Internet Search', mimeo

Schaefer, M. and G. Sapi (2022) 'Complementarities in Learning from Data: Insights from General Search', mimeo, available at https://dx.doi.org/10.2139/ssrn.4460899

Scott-Morton, F. and D. Dinielli (2020) *Roadmap for a Digital Advertising Monopolization Case Against Google*, Omidyar Network

Sevilla, J., L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn and P. Villalobos (2022) 'Compute Trends Across Three Eras of Machine Learning', mimeo, available at https://arxiv.org/abs/2202.05924

US House of Representatives (2020) *Investigation of competition in digital markets*

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin (2017) 'Attention Is All You Need', mimeo, available at https://arxiv.org/abs/1706.03762

Wu, Z., M. Sanderson, B. Cambazoglu, W.B. Croft and F. Scholer (2020) 'Providing Direct Answers in Search Results: A Study of User Behavior', *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*: 1635-44

Zhao, X., S. Ouyang, Z. Yu, M. Wu and L. Li (2022) 'Pretrained language models can be fully zero-shot learners', mimeo, available at https://arxiv.org/abs/2212.06950

Bruegel, Rue de la Charité 33, B-1210 Brussels
(+32) 2 227 4210
info@bruegel.org
www.bruegel.org