PIIE PETERSON INSTITUTE FOR INTERNATIONAL ECONOMICS

POLICY BRIEF

19-3 Opening Internet Monopolies to Competition with Data Sharing Mandates

Claudia Biancotti and Paolo Ciocca

April 2019

Claudia Biancotti is visiting fellow at the Peterson Institute for International Economics and senior economist at the Bank of Italy. **Paolo Ciocca** is commissioner at Consob, the Italian stock exchange commission.

Authors' Note: We are grateful to Martin Chorzempa, Mario Rasetti, Edwin Truman, and participants at a PIIE internal seminar for their helpful comments and suggestions. Any remaining errors or omissions are solely our responsibility. The views expressed here are our own and should not be attributed to PIIE, its Board of Directors, the Bank of Italy, or Consob. This Brief extends the analysis presented in an earlier column (Biancotti and Ciocca 2018).

Hardly anyone is unaware of the revolutionary new avenues of commerce, communication, and information sharing dominated by a small number of technology companies mobilizing datasets of unprecedented size and granularity. The five biggest of these tech behemoths—Google, Apple, Facebook, Amazon, and Microsoft (GAFAM to the cognoscenti)—have risen to become the richest and most pervasive companies in the world, vacuuming up the characteristics, preferences, and behavior of billions of individuals. By extracting progressively more accurate insight from this data via artificial intelligence (AI) tools, such as machine learning (ML), internet firms can offer consumers highly customized experiences. This has no doubt made communicating, sharing, searching, and doing business more convenient for everyone. An increasingly popular term for this brave new world is "surveillance capitalism."

But these concentrations of data are at the root of a worrying power imbalance between dominant digital companies (GAFAM especially) and the rest of society. Despite recent advances in privacy legislation, users still have little control over who sees the information they provide and how it is monetized—that is, exploited by these five companies to generate hundreds of billions of dollars in revenue from targeted advertising and other services that leverage user information. As exclusive gateways to information for many, the GAFAM companies are positioned to stifle competition without protecting user privacy. Controls that might prevent personalized content from being used by hostile actors to influence public debate are lax or nonexistent, as is now well known in the United States, where discussion has become polluted by foreign entities and hate groups, in some cases advocating violence.

This Policy Brief outlines the range of issues posed by the dominance of the GAFAM companies, citing several concerns in the areas of collective security, consumer rights, and competition—and offers a possible solution to the latter.

To begin with the security issue, data breaches at Facebook and elsewhere have made it painfully obvious that the data stored by these companies is not always secure and can be stolen by malign actors. Second, consumer rights to privacy regarding online behavior are an increasing concern in Europe but also America. Third is the problem of competition—the concern that the dominance of companies possessing exclusive information on a majority of internet users and a superior ability to recruit analytical talent allows them to prevent market entry of potential competitors who might ultimately deliver better goods and services or foreclose the existing ones before they grow big enough to pose a threat.

Solutions to these problems are complicated because policymakers are dealing in a world of tradeoffs, in which individuals often willingly surrender their privacy to take advantage of what these companies offer. It may thus be hard to regulate voluntary behavior—protecting people's liberties as well as their privacy. There are even tradeoffs suggesting that moving on one front jeopardizes the quest for progress on other fronts.

This Policy Brief argues, for example, that one promising idea dealing with barriers to competition would be to introduce data sharing mandates, or requirements for market leaders to share anonymous user data with competitors and with academia. AI algorithms need to be trained on large datasets to work properly, and entrants in the growing number of markets where AI is crucial to success face the so-called "cold start problem." They have no users yet, which means they have no data. It is very hard for them to compete with incumbents who have indepth knowledge of their existing users and can readily project it onto newcomers.

Data sharing mandates would help solve this problem. Data- and AI-driven markets would become more competitive, and the benefits of AI would become more widespread throughout the economy. On the other hand, such a measure could worsen existing risks to consumer privacy and collective security, because at present there are no methods to ensure the anonymity of highly granular datasets. This Policy Brief concludes by arguing that policymakers intending to implement a data sharing mandate should carefully evaluate the tradeoff between the comprehensiveness of the information shared and re-identification risks, while also providing new incentives for research on anonymization techniques.

WHY DATA—IN UNPRECEDENTED VOLUME—MATTERS

In 2009, a group of computer scientists from Princeton University published ImageNet, a dataset that initially contained 3.2 million digital photographs representing approximately 5,000 real-world objects (Deng et al. 2009). This marked a turning point in computer vision, as machines finally had enough data to learn from to understand what different objects look like. By 2015, computers started outperforming humans in object recognition tasks (He et al. 2016).

ImageNet was the opening act for the boom in machine learning (ML), a type of artificial intelligence (AI) based on algorithms that derive decision rules from observed examples. ML models require large quantities of data. For decades, this constituted a barrier to their adoption, despite the fact that the mathematical foundations had been laid as early as the 1950s (Rosenblatt 1958). As the internet expanded and economies digitized, researchers and companies had more and more information to work with, and ML became commonplace (Agrawal, Gans, and Goldfarb 2018; Cockburn, Henderson, and Stern 2018).

Over time, AI applications evolved from simple tasks such as classifying static objects to considerably more complex endeavors, which further increased—and will continue to increase—the relevance of data for technical, social, and economic progress. For example, digital images from brain scans are a key input in semantic mapping, a technology that promises to improve the quality of life for those who suffer from conditions that limit the ability to speak (Huth et al. 2016). The real-time analysis of video feeds and other information collected from onboard sensors is what enables autonomous vehicles to avoid collisions. Massive datasets of annotated legal text teach machines how to review contracts, improving efficiency and reducing error margins in interpretation.[1] In finance, algorithms are expanding beyond the trading floor into retail applications, such as personalized investment advice. The Organization for Economic Cooperation and Development (OECD) has defined data-driven innovation as "a key pillar of 21st century economic growth" (OECD 2015).

The GAFAM companies enjoy a significant data advantage over competitors. In 2017, Google reported that its Android operating system was installed on more than 2 billion devices active at least once a month.[2] In 2018, Facebook's flagship social platform had 2.3 billion users,[3] nearly 60 percent of the global population with internet access.[4] For comparison, Twitter and Reddit, among the few popular platforms not acquired by GAFAM companies, hovered around 330 million users each.[5]

Although the GAFAM companies differ in terms of the exact set of variables they collect, each of them knows its users in most of the following dimensions: personally identifying information, including physical characteristics; social contacts; geographical location; employment; beliefs, opinions, and preferences; and actions performed while online, which may include web pages visited, products bought, amount of money spent, links clicked, videos watched, and searches conducted. Data collection on offline activities, such

[1] Hugh Son, "JPMorgan Software Does in Seconds what Took Lawyers 360,000 Hours," Bloomberg, February 27, 2017, https://www.bloomberg.com/news/articles/2017-02-28/jpmorgan-marshals-an-army-of-developers-to-automate-high-finance.

[2] Ben Popper, "Google Announces Over 2 Billion Monthly Active Devices on Android," The Verge, May 2017, https://www.theverge.com/2017/5/17/15654454/android-reaches-2-billion-monthly-active-users.

[3] Monthly active users. "Facebook Reports Third Quarter 2018 Results," Facebook press release, October 30, 2018, https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-Third-Quarter-2018-Results/default.aspx.

[4] Projected number of internet users by the end of 2018. See International Telecommunication Union, ICT Statistics, https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.

[5] "Twitter Announces Third Quarter 2018 Results," Twitter press release, October 25, 2018, https://s22.q4cdn.com/826641620/files/doc_news/events/2018/Q3-2018-Earnings-Press-Release.pdf. Reddit by the Numbers, available at https://www.redditinc.com/press (accessed on March 18, 2019).

as credit card use in physical stores, in combination with online behavior is increasingly common.[6] This represents a unique window into human behavior.

THE RISKS OF DATA CONCENTRATION

The first set of risks associated with data concentration concerns collective security.

Technical vulnerabilities are inherent to all computer systems, and the crowd of hostile actors willing to launch cyberattacks that exploit them for profit or to achieve strategic aims is large. The GAFAM companies are a prime target on both counts because of the data they own and the input they provide for other economic activities.

Large breaches of GAFAM data would have fallout that goes beyond the sum of perceived privacy violations. An individual who takes a picture of a landscape and posts it online may consider it less private compared to a family photo and not worry excessively if hackers access it. The picture, however, still embeds personal information such as the photographer's location at a given time and date. Attackers, when looking at the location of many individuals via ML techniques, can infer information on sensitive matters such as military operations[7] or political initiatives.

When it comes to hacks aimed at business disruption, again, damage exceeds losses borne by the immediate victim. The negative externalities of cyberattacks are evident (Anderson and Moore 2011). Should the services of a leading cloud computing provider—e.g., Amazon Web Services—be unavailable even for a short time, the consequences would extend to all firms relying on it and to their customers. The same applies for other GAFAM products that are widely used across the economy.

Attacks have happened in the past. Facebook has suffered multiple data breaches, the latest affecting about 30 million users.[8] In November 2018, Google internet traffic was briefly diverted to China:[9] The company maintains that this was because of human error, yet such diversions have been shown to happen in other cases—not involving Google—as a result of deliberate actions on the part of China Telecom (Demchak and Shavit 2018). None of these episodes were of striking proportions, but the threat is evident.

On this front, the interests of the GAFAM companies and those of society are aligned: Both want to avoid attacks. While some implementation questions remain unanswered, for example, how best to allocate cybersecurity investment between the public and private sectors, there is no fundamental rift.

The picture is more complex when it comes to another link between data and collective security, i.e., the ability of hostile actors to leverage the near-universal reach of the GAFAM companies and harness their ML-based insights to manipulate public opinion. In this case, public and private interests can be at odds, with corporations wanting to sell advertising space to the highest bidder and regulators wanting to protect the integrity of the democratic process.

So far, the industry has largely been left to self-regulate. Especially in the United States, lawmakers have been giving mixed signals: They have heard allegations, in testimony and in the public discourse, of GAFAM companies facilitating the manipulation of information and even taking an active role in the process for their own ends,[10] but they have not put forward any policy responses. The problem is difficult, not least because the line between legitimate attempts at persuading others—including in the political arena—and maliciously distorting information is hard to draw in a way that makes legal sense in the context of online platforms. Moreover, laws aimed at controlling published content tend to raise censorship concerns.[11]

As more evidence of data-driven misinformation emerges,[12] however, inaction could be costly both to society

[6] Mark Bergen and Jennifer Surane, "Google and Mastercard Cut a Secret Ad Deal to Track Retail Sales," Bloomberg, August 30, 2018, https://www.bloomberg.com/news/articles/2018-08-30/google-and-mastercard-cut-a-secret-ad-deal-to-track-retail-sales. Amazon offers credit cards in partnership with Visa.

[7] Alex Hern, "Fitness Tracking App Strava Gives Away Location of Secret US Army Bases," *The Guardian*, January 28, 2018, https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases.

[8] Salvador Rodriguez, "Facebook Says Hackers Were Able to Access Millions of Phone Numbers and Email Addresses," CNBC.com, October 12, 2018, https://www.cnbc.com/2018/10/12/facebook-security-breach-details.html.

[9] Dan Goodin, "Google Goes Down After Major BGP Mishap Routes Traffic Through China," Ars Technica, November 13, 2018, https://arstechnica.com/information-technology/2018/11/major-bgp-mishap-takes-down-google-as-traffic-improperly-travels-to-china/.

[10] See, for example, the December 10, 2018 testimony of Alphabet CEO Sundar Pichai at the US House of Representatives.

[11] Adam Minter, "Fake News Laws are Fake Solution," Bloomberg Opinion, May 24, 2018, https://www.bloomberg.com/opinion/articles/2018-05-25/fake-news-laws-are-fake-solution.

[12] UK House of Commons—Digital Culture, Media and Sport Committee, "Disinformation and 'fake news': Final Report," February 2109, https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf. European Commission, "A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation," March 2018, https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation.

and to GAFAM companies themselves. There may be a tipping point after which public trust in internet platforms breaks down. Suspicion could spread to other digital environments and trigger abandonment of at least some new technologies, with the attendant harm to productivity and growth.

The second risk associated with data concentration concerns consumer rights.

A growing body of literature in economics and law is devoted to understanding whether data concentration may constitute barriers to entry in certain markets.

Some issues that arise when firms acquire and use data provided by and related to individuals are covered by consumer protection laws. In these cases, size is not relevant—all data collectors are affected, independent of how many customers they have. In the many grey areas that still exist, however, the GAFAM companies are implicitly setting standards, since their decisions are affecting billions of people.

The one field where some consensus has been achieved is privacy, defined as an individual's ability to separate the private from the public sphere by setting limits on who can access specific pieces of information (Acquisti, Taylor, and Wagman 2016). Most OECD countries either have legislated on the matter or are in the process of doing so, building on the idea that the use, sharing, and selling of personal data collected by firms should require consent on the part of the person. While significant doubts exist on the effectiveness of this approach,[13] and the United States still does not have federal privacy laws, some progress is being made.

Conversely, the asymmetry in power and information between individuals who provide data and companies that extract revenue from it has not been addressed extensively. The GAFAM standard is provision of digital services free of charge in exchange for user data, but economists have questioned whether these terms fairly reflect the value of the information (Arrieta Ibarra et al. 2018).

The existence of for-profit brokers of personal data such as Acxiom, Experian, and Bluekai—some of whom even sold data to GAFAM companies[14]—shows that an information market with monetary rewards exists. So far, however, individuals have not been involved directly and they have not been able to appropriate any significant share of the value. When it comes to ML, data are profitable only in bulk: The marginal return of a single record is negligible. As long as data subjects try to sell their information individually, their bargaining power is minimal. Payments offered by consumer-oriented data monetization apps are very small.[15]

Academia and advocacy groups have explored a few possible solutions. Lanier and Weyl (2018) suggest that individuals should coordinate via union-like organizations that would negotiate with corporations on their behalf. Technical means for consumers to track who has access to their information are also being developed, as exemplified by the Solid Project at the Massachusetts Institute of Technology.[16] Increased awareness on data circulation would also reduce asymmetries between service providers and users. At the beginning of 2019, the idea that consumers should share in the value created through their data appeared to be gaining some traction in mainstream US political discourse.[17]

Data-driven discrimination is another area that needs attention. In 2018, the US Department of Housing and Urban Development found that Facebook offered landlords and property developers the possibility to advertise selectively based on race. Facebook has since pledged to stop this practice,[18] but the problem is far from being solved. Regulators have difficulties detecting subtler forms of algorithmic bias, where variables such as ethnicity, gender, age, and religion may be improperly factored into decisions, and finding proof is difficult because the variables are lost amidst other factors in black-box models (Boddington 2017).

[13] See for example Schaub, Balebako, and Cranor (2017) on the difficulties that average internet users have in reading and understanding privacy policies written in complex legal language.

[14] Sonam Rai, "Acxiom Shares Tank After Facebook Cuts Ties with Data Brokers," Reuters, March 29, 2018, https://uk.reuters.com/article/us-acxiom-stocks/acxiom-shares-tank-after-facebook-cuts-ties-with-data-brokers-idUKKBN1H520U.

[15] Gregory Barber, "I Sold My Data for Crypto. Here's How Much I Made," *Wired*, December 17, 2018, https://www.wired.com/story/i-sold-my-data-for-crypto/.

[16] The Solid Project is located at https://solid.mit.edu/.

[17] Kartikay Mehrotra, "California Governor proposed Digital Dividend Aimed at Big Tech," Bloomberg, February 12, 2019, https://www.bloomberg.com/news/articles/2019-02-12/california-governor-proposes-digital-dividend-targeting-big-tech.

[18] Nick Statt, "Facebook will Remove 5,000 Ad Targeting Categories to Prevent Discrimination," The Verge, August 21, 2018, https://www.theverge.com/2018/8/21/17764480/facebook-ad-targeting-options-removal-housing-racial-discrimination.

Types of discrimination that are not subject to general prohibitions, such as differential pricing based on a consumer's characteristics and estimated willingness to pay, still raise legal and ethical issues when performed on a very large scale and/or in sensitive sectors (Executive Office of the President of the United States 2014). In insurance, moves toward highly personalized risk assessment and price discrimination reduce the benefits of mutualization for insureds evaluated as high risk, while preserving the benefits of risk-pooling for insurers and increasing their profits.

Finally, data concentration poses a risk to competition.

The most evident channel through which data gives GAFAM companies a competitive advantage is the ability to obtain better predictions from ML algorithms. This has an impact on:

- Flagship product markets, e.g., internet search for Google and ecommerce for Amazon. The GAFAM companies command large market shares for their most popular offerings. In some cases, this already borders on monopoly.[19] In other instances, market share is not as large, but the leader still greatly outpaces the others.[20] Being so far ahead of the rest means collecting a lot more information on what individuals do, which in turn allows these companies to improve the user experience via ML more quickly than others on a continuous basis.
- Other product markets. The GAFAM companies develop a variety of products that, at least in part, leverage the same data and technologies for different purposes. For example, text typed in search queries or emails can be used to improve natural language processing capabilities for digital home assistants. The combination of network effects and data advantage conferred by the flagship offering can be used to gain dominant positions in other areas. Cross-product externalities can be especially large for multiple services consumed from a single account in a closed, data-intensive environment such as a smartphone, but they extend to other domains including the physical.[21]

The GAFAM competitive advantage extends to technology development, a fact not often noticed but which is more relevant in the long run. ML is an example of narrow artificial intelligence (NAI), or AI whose abilities are confined to specialized tasks. The next step, still to be attained, is machine reasoning capable of human-like creativity and flexibility (artificial general intelligence, AGI, also referred to as strong AI). Computer scientists disagree over how far off in the future AGI is[22] and how gradually NAI will progress to AGI.[23]

In any case, more data are likely to help the GAFAM companies develop AGI faster than competitors. One possible approach to AI research is getting a computer to mimic human behavior: These companies are particularly well-placed to understand human decision making, because they have a constant live stream of information on what billions of people choose to do. They are also able to attract scientific talent—a necessary complement to data—on par with top research institutions.[24]

If the GAFAM companies made significant progress toward AGI before everyone else, their competitive lead in flagship and other markets would be substantially reinforced—they would be the first to deploy new, more sophisticated ways to attract and retain users, and to extract profits from them. As AI research is moving from universities and the public domain to the private sector and patents, GAFAM gains from AGI can be seen as a form of vertical integration.

Economic theory posits that the existence of competitive advantages, either in product markets or research, is not negative *per se*. The prospect of enjoying market power and

[19] In 2018 nearly 90 percent of internet searches in the United States went through Google Search, and over 70 percent of desktop computers ran Microsoft Windows (StatCounter GlobalStats, http://gs.statcounter.com/).

[20] Amazon controls a little less than half of total US ecommerce, with eBay coming in a distant second at under 7 percent (Ingrid Lunden, "Amazon's Share of the US e-Commerce Market is now 49 percent, or 5 percent of all Retail Spend," TechCrunch, August 2018).

[21] Amazon's expansion into health care provides a clear illustration, as the company jointly leverages its leadership in online marketplaces, cloud computing infrastructure, and digital home assistants to make a strong entrance in hospital supply chains, the retail pharmaceutical market, medical research, and care services. See Alia Paavola, "Amazon Moves into Healthcare: A 2018 Timeline," Becker's Hospital Review, December 20, 2018, https://www.beckershospitalreview.com/healthcare-information-technology/amazon-moves-into-healthcare-a-2018-timeline.html.

[22] Emerging Technology from the ArXiv, "Experts Predict When Artificial Intelligence Will Exceed Human Performance," *MIT Technology Review*, May 31, 2017, https://www.technologyreview.com/s/607970/experts-predict-when-artificial-intelligence-will-exceed-human-performance/.

[23] AI pioneer Judea Pearl (2018) recently argued that "[ML] systems cannot reason about interventions and retrospection and, therefore, cannot serve as the basis for strong AI." Conversely Yann LeCun, who first introduced convolutional neural networks, thus argued on Quora: "I personally believe that there cannot be breakthroughs towards AI that do not make strong use of machine learning. Machine learning is where it's at, and representation learning (or deep learning) is where the action is." Thread available at https://www.quora.com/What-AI-breakthroughs-have-we-had-towards-strong-AI-outside-of-Machine-Learning.

[24] Kaveh Waddell, "A Feud Atop AI's Commanding Heights," Axios, September 6, 2018, https://www.axios.com/academia-corporate-research-ai-9d525070-303d-47fd-b822-0fbffcac6740.html.

reaping profits for some time is what incentivizes firms to innovate. The problem arises when dominant firms foreclose potential competitors. In the case of the GAFAM companies, this may happen first in flagship markets, with other ones following as a consequence. Entrenched dominance can lead to permanent supracompetitive prices and/or degraded quality for consumers. It may also result in less innovation, since the dominant firm can enjoy rents without having to invest, although evidence on whether this argument applies to the GAFAM companies is mixed.[25]

A growing body of literature in economics and law is devoted to understanding whether data concentration may constitute barriers to entry in certain markets.[26] Such barriers appear to exist in the case of GAFAM flagship markets, given the combination of network externalities, increasing returns to scale, lock-in effects, and ML-induced feedback loops whereby incumbents can improve their services more quickly than newcomers because they have access to more data.[27]

This advantage spills over into adjacent markets, where indeed EU antitrust enforcers have ascertained exclusionary conduct. Between 2017 and 2018, the European Commission levied large fines on Google for abusing its dominance in internet search[28] and mobile operating systems[29] to restrict competition in other digital markets. With respect to AI development, the combination of unique insights on behavior and talent advantage poses similar risks.

Monopolization of markets that are farther away from core internet-based services appears to be somewhat less likely right now, as many competitors that have accumulated significant domain-specific expertise and data over time exist. Technological change, however, is likely to tip the balance in the favor of the GAFAM companies. In the market for driverless cars, long-standing automotive leaders are struggling to keep up with Google's Waymo.[30] Amazon's expansion into logistics provides another example.

DATA SHARING MANDATES

In August 2018, Andrea Nahles, leader of the Social Democratic Party of Germany, advocated for legislation that would require digital companies above a certain size to share a representative slice of their user data with the public. The sharing, she argued, would open new opportunities for smaller companies, reducing inequalities and fostering growth.[31]

A few months earlier, *The Economist* had called for a slightly different scheme,[32] suggesting that tech market leaders should give access to some of their user data to competitors in exchange for a fee. In an essay published in *Foreign Affairs,* Oxford internet governance expert Viktor Mayer-Schönberger and technology journalist Thomas Ramge[33] put forward a progressive version: "[E]very company above a

[25] On the one hand, the emergence of monolithic corporate cultures may indeed reduce the drive to innovate. Upon leaving Facebook, founders of high-profile apps such as WhatsApp and Instagram cited an excessive focus on profits from advertising as opposed to delivering better products (Tim Bradshaw and Aliya Ram, "Instagram Founders Quit Facebook-Owned Photo App," *Financial Times*, September 25, 2018, https://www.ft.com/content/0afda1ae-c070-11e8-95b1-d36dfef1b89a). On the other hand, the GAFAM companies have a track record of consistently investing in research and are responsible for an increasing number of advancements: Indeed, the ImageNet breakthrough came from Microsoft. As the market for internet-based services globalizes, it is also increasingly likely that the GAFAM companies will keep on innovating as they start to feel competitive pressure from Chinese firms, independent of how large their OECD market share is right now.

[26] See Autorité de la Concurrence and Bundeskartellamt 2016; Bourreau, de Streel, and Graef 2016; and Rubinfeld and Gal 2017 for a comprehensive overview.

[27] Two objections are frequently raised to this claim. One is based on the rapid ascent of Google and Facebook in 1998 and 2004 respectively. The two companies were able to gain dominance on the sole merit of their good ideas, unseating powerful incumbents Yahoo! Search and Myspace. However, this was long before ImageNet, and ML did not have the same relevance; having more data did not substantially favor incumbents. A different argument is based on the fact that ML has decreasing returns to scale, as shown by many experiments, and therefore the marginal contribution of the billionth record is irrelevant. Posner and Weyl (2018) show that this point is only partially valid, as the same set of data can be employed to perform different tasks, each with a different sample complexity or minimum sample size needed to achieve the goal. Decreasing returns to scale can exist within each task, although this is not guaranteed. Over a continuum of progressively more difficult tasks, for example, from recognizing the presence of humans in a photograph to putting a label on the action they are performing, returns to scale can well be increasing.

[28] European Commission, "Antitrust: Commission fines Google €2.42 billion for abusing dominance as search engine by giving illegal advantage to own comparison shopping service," press release, June 27, 2017, http://europa.eu/rapid/press-release_IP-17-1784_en.htm.

[29] European Commission, "Antitrust: Commission fines Google €4.34 billion for illegal practices regarding Android mobile devices to strengthen dominance of Google's search engine," press release, June 18, 2017, http://europa.eu/rapid/press-release_IP-18-4581_en.htm.

[30] David Welch and Elisabeth Behrmann, "Who's Winning the Self-Driving Car Race?" *Bloomberg*, May 7, 2018.

[31] Andrea Nahles, "Die Tech-Riesen des Silicon Valleys gefährden den fairen Wettbewerb," *Handelsblatt*, August 13, 2018, https://www.handelsblatt.com/meinung/gastbeitraege/gastkommentar-die-tech-riesen-des-silicon-valleys-gefaehrden-den-fairen-wettbewerb/22900656.html.

[32] "How to Tame the Tech Titans," *The Economist,* January 12, 2018.

[33] Viktor Mayer-Schönberger and Thomas Ramge, "A Big Choice for Big Tech," *Foreign Affairs*, September/October 2018.

certain size […] that systematically collects and analyzes data would have to let other companies in the same market access a subset of its data. The larger a firm's market share, the more of its data others would be allowed to see."

Calls for data sharing mandates (DSMs) partly reflect, in a novel and more comprehensive way, long-standing concerns in competition policy about individual companies monopolizing key inputs. In 2008, the US Department of Justice (DoJ) approved the merger of financial data providers Thomson Corporation and Reuters Group, conditional on Thomson selling copies of three proprietary datasets and licensing related intellectual property to a firm or firms that would use the data to offer products and services in competition with the merged entity. Exclusive ownership of the datasets on the part of a single company, the DoJ argued, "likely would have led to higher prices and reduced innovation."[34] The EU Commission reached a similar conclusion.[35]

In the Thomson Reuters merger, the competitive relevance of the data was straightforward, because the merging parties were in the business of selling it. Competition authorities have been criticized in the literature (see e.g., Chirita forthcoming 2019) for failing to appreciate, in the very first years after the ImageNet breakthrough, the potential anticompetitive effects of data concentration in cases relating to firms that do not sell information but rather use it as an input for other products. Today, this profile is starting to be considered routinely in merger reviews in both the European Union and the United States.[36] In early 2019, it was a crucial factor for the first time in an antitrust decision involving an internet giant, as the German competition authority prohibited Facebook from linking user data across different services absent user consent[37] and clearly framed the practice as an abuse of dominant position.

A DSM that obligates companies to sell data to competitors, as opposed to giving it away at no charge, could also be seen as an application of the essential facilities doctrine—which posits that a firm with exclusive control over a facility that is essential for other firms to effectively compete in a downstream market has an obligation to grant access to it in exchange for a reasonable price. The object of controversy among legal scholars, this doctrine was nonetheless a factor in decisions by antitrust courts both in the United States and the European Union (Pitofsky, Patterson, and Hooks 2002; Graef 2017).

Among the many possible measures that could address the negative effects of data concentration, DSMs appear particularly interesting in that they aim at widening the set of economic agents that can extract value from information without unduly constraining incumbents. In all matters of competition, policymakers must strike a delicate balance. They must fight abuses of dominant position and prevent complacency on the part of market leaders, making sure that market entry on fair terms is possible and that continued innovation is needed to maintain an edge. At the same time, they must avoid interventions that feel like arbitrary redistribution of profits that could discourage both leaders and newcomers from investing.

When it comes to the GAFAM companies, finding this balance is especially important. Given their capacity for innovation and their role of quasi-infrastructure for the digital economy, the wrong incentive mix could cripple the overall rate of technological progress, ultimately hurting growth. Moreover, the market for internet-based services is starting to globalize, but while some jurisdictions are open to giving foreign providers market access, others place restrictions on it. Any US or EU policy intervention or enforcement strategy aimed at containing the negative effects of GAFAM dominance should take into account the absence of a level playing field. The current asymmetry in access should not favor players originating in closed markets.

POLICY TRADEOFFS

If evaluated exclusively on their potential competition merits, DSMs appear fundamentally good, albeit somewhat limited in reach. They would afford more companies the possibility to derive ML-based insights—or improve their existing ones—in a wide variety of fields dependent on analyzing human behavior. With respect to the GAFAM companies flagship markets, and adjacent markets established within cohesive digital ecosystems, data sharing would partially remove barriers to entry, although it may not be sufficient on its own, given the network externalities and lock-in effects enjoyed by incumbents.

With respect to markets where the GAFAM companies are expanding but are not dominant yet, a DSM would reduce risks of future monopolization by allowing productivity gains from ML to spread throughout the economy. This would especially benefit small and medium enterprises (SMEs) that are now unable to employ ML because they do not have enough information. All business owners would be able to obtain information on consumer opinions and decisions relating to

[34] US Department of Justice, "Justice Department Requires Thomson to Sell Financial Data and Related Assets in Order to Acquire Reuters," press release, February 19, 2008.

[35] European Commission, "Mergers: Commission Clears Acquisition of Reuters by Thomson Subject to Conditions," press release, February 19, 2008.

[36] See for a recent example: European Commission, "Mergers: Commission Clears Apple's Acquisition of Shazam," press release, September 6, 2018.

[37] Bundeskartellamt, "Bundeskartellamt prohibits Facebook from combining user data from different sources," February 7, 2019, https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2019/07_02_2019_Facebook.html.

their sector of activity, competitors, and products. Design, production, and marketing choices would improve. So would overall economic performance.

This opportunity might be somewhat curbed by the fact that analytics skills are not very common in SMEs. Wider data availability, however, could also lead to higher competitiveness and lower prices in the market for analytics as a service: Right now, specialized firms sell a mixture of private information and ability to generate insights, but under a DSM the degree of market power derived from private information would be reduced.

DSMs may also benefit AI development by fostering competition. They could multiply the opportunities for improving on existing ML algorithms by giving more researchers the possibility to experiment with different specifications and accelerate progress toward AGI by distributing knowledge on human decision making. Similarly, there could be an upside for reflections on AI ethics, as the number of people with firsthand experience of possible distorted uses of large datasets increases across national and corporate cultures.

Such progress is not guaranteed because the AI research talent pool is still very small, with some estimates putting the global number of top-flight experts at about 22,000.[38] While the number may be overly pessimistic, research skills are certainly much less common than the ability to use off-the-shelf ML tools in business. The majority of non-GAFAM companies do not have the means to attract top talent, so they would not be able to compete effectively even if they had much more data. One way around this problem would be to specify in DSMs that academic institutions, which still employ some experts, be included among data recipients. This would both expand research opportunities in absolute terms and shift some weight back to the nonprofit, non-applied world.

Beyond the issue of competition, though, DSMs are not an unambiguously positive solution. One key problem is privacy protection. In several jurisdictions, including the European Union and California, it is illegal for companies to share the personal data they collect with third parties without the consent of the data subject. It is legal to share anonymized data. Definitions of anonymization vary across countries, but the common principle is that information should be stripped of all characteristics that make it attributable to a specific individual. Assuming legislators would not care to sacrifice popular data protection statutes for the sake of facilitating data sharing, any DSM would entail either consent from data subjects to share identifiable information or data anonymization.

Relying on consent is highly problematic. Preferences for privacy vary widely, and are difficult to measure and sometimes contradictory. A growing literature discusses the so-called privacy paradox (for a review, see Kokolakis 2017). Many individuals declare that they value their privacy highly, while regularly revealing information online in exchange for small rewards such as “likes” on social platforms. Determinants of this paradox are not well understood yet. If consent is a requirement, user self-selection would introduce a bias in the shared data. The direction of the bias would be unknown to the beneficiaries of the DSM, while the GAFAM companies could estimate it and bank on this further advantage.

Multiplying the number of entities that can access personal information also increases opportunities for discrimination and exposure to cyberattacks, as the protection standards of smaller companies are generally worse compared to larger ones (Biancotti 2017). In turn, this threatens both user privacy and collective security, even if all parties to the sharing are in compliance with data protection laws.

Moreover, a requirement that personal identifiers be shared might make DSMs too costly for the GAFAM companies. Their revenues come from two sources. One is the ability to develop products that consumers like; the other is the availability of a large audience for marketing purposes. They feed back into each other, but they are not the same. The former does not need consumer identification; the latter depends on it.

ML algorithms aimed at, say, predicting which new music an individual will like do not need to know the person’s name. Anonymous information on demographics, location, and listening habits is sufficient. Advertisers wanting to promote records to that specific individual, conversely, need to have personal access to him or her. A DSM that does not include identifying information fosters competition by allowing entrants to leverage ML, but it does not immediately destroy the exclusive marketing reach of the GAFAM companies.

For these reasons, DSMs should require that data are anonymized before they are shared, but it is important to note that this solution still falls far short of eliminating risk. Anonymization methods that have proven successful for traditional statistical surveys—such as outlier deletion and limited data obfuscation—do not perform well in the new world of huge, complex, highly granular datasets (Bohannon 2015, Torra and Navarro-Arribas 2016). Such datasets are particularly vulnerable to re-identifing data subjects through a variety of techniques.[39]

[38] Jeremy Kahn, “Just How Shallow is the Artificial Intelligence Talent Pool?,” Bloomberg, February 7, 2018, https://www.bloomberg.com/news/articles/2018-02-07/just-how-shallow-is-the-artificial-intelligence-talent-pool.

[39] For example, in what is known as a linkage attack, a sufficiently large number of features unique to an individual re-

Research is making progress on how to better defend against this possibility and generate privacy-preserving synthetic data. Experiments have shown that, given enough real-world information, machines can generate simulated data that have no connection to any existing individual but are equally effective in training ML models (Patki, Wedge, and Veeramachaneni 2016; Aviñó, Ruffini, and Gavaldà 2018). Indeed, DSMs could accelerate progress in this field, by giving more companies a starting point to create their own synthetic datasets and to develop novel methods to do so.[40] In turn, this could go some way towards what has been called the "democratization" of ML,[41] by making companies less dependent on large user bases for their ML endeavors.

This is, however, still some way off in the future. Until new anonymization methods consistently deliver good results, certain categories of information—such as the physical locations that some apps record—would have to be excluded from a DSM because there is no effective way of anonymizing them. Other data would have to be heavily edited. Any DSM based on anonymization implies a tradeoff between privacy protection and the economic value of the information, which decreases after deletions and manipulations. This tradeoff needs to be carefully evaluated. Doing so requires expanding the very limited empirical knowledge of which pieces of information are most valuable economically.

Competition-privacy tradeoffs are not the only ones that may arise when designing policies for the redress of power imbalances in the digital economy. Fostering the creation of data markets, where returns from information are shared more equally between consumers and corporations, is a worthy goal in principle. It may be achieved through appropriate specifications of DSMs or through different measures.

On the other hand, security risks could emerge. If individuals could profit from selling their own data, the amount of information disclosed would likely increase and so would the number of parties that have access to it. The GAFAM companies would not be the only potential buyers in an open data market: Any company that employs ML or wants to expand its reach would show an interest. Safeguards would be needed to prevent hostile actors from posing as legitimate businesses and buying personal data in bulk.

CONCLUSIONS

Over the past few years, it has become apparent that a small number of technology companies have assembled detailed datasets on the characteristics, preferences, and behavior of billions of individuals. This concentration of data is at the root of a worrying power imbalance between dominant internet firms and the rest of society, reflecting negatively on collective security, consumer rights, and competition.

Introducing data sharing mandates, or requirements for market leaders to share anonymous user data with other firms and academia, would have a positive effect on competition. As data are a key input for AI, both in terms of applications and research, more widely available information would help spread the benefits of AI through the economy. On the other hand, data sharing could worsen existing risks to consumer privacy and collective security, because at present there are no failsafe methods to ensure the anonymity of highly granular datasets.

Policymakers intending to implement a data sharing mandate should carefully evaluate this tradeoff when defining exactly which pieces of information must be shared by dominant firms. As the mandate becomes more comprehensive, data- and AI-driven markets will become more competitive, while re-identification risks for data subjects will worsen. Data sharing mandates may have to start relatively small, only encompassing information associated with minimal probability of reconstructing individual identities. They should be accompanied by new incentives for research on anonymization techniques, which would allow them to grow in scope and effectiveness over time.

cord is extracted from an anonymized dataset and matched to other information sets, which may include public social media footprints, where personal identifiers are available (Garfinkel 2015).

[40] In subfields of ML such as computer vision, where many features of the data can be accurately simulated based on mathematical models—say, changes in an object's appearance that follow changes in lighting or motion speed—synthetic information may eventually be generated with almost no need for real-life examples (see Mayer et al. 2018). Such a development is unlikely in fields that have to do with human behavior, which is why the availability of real data as a starting point will remain relevant.

[41] Bernard Marr, "Does Synthetic Data Hold the Secret to Artificial Intelligence?," *Forbes*, November 5, 2018, https://www.forbes.com/sites/bernardmarr/2018/11/05/does-synthetic-data-hold-the-secret-to-artificial-intelligence/#3b0299ed42f8.

REFERENCES

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman. 2016. The Economics of Privacy. *Journal of Economic Literature* 54(2): 442–92.

Agrawal, Ajay, Avi Goldfarb, and Joshua Gans. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Cambridge: Harvard Business Review Press.

Anderson, Ross, and Tyler Moore. 2011. Internet Security. In *The Oxford Handbook of the Digital Economy*, ed. Martin Peitz and Joel Waldfogel. Oxford: Oxford University Press.

Arrieta Ibarra, Imanol, Leonard Goff, Diego Jiménez Hérnandez, Jaron Lanier, and E. Glen Weyl. 2018. Should We Treat Data as Labor? Moving beyond "Free." *American Economic Association Papers and Proceedings* 108(1): 38–42.

Autorité de la Concurrence and Bundeskartellamt. 2016. *Competition Law and Data*. Paris and Bonn.

Aviñó, Laura, Matteo Ruffini, and Ricard Gavaldà. 2018. Generating Synthetic but Plausible Healthcare Record Datasets. *Proceedings of 2018 KDD workshop on Machine Learning for Medicine and Healthcare.* 2018 Knowledge Discovery and Data Mining Conference, London.

Biancotti, Claudia. 2017. The Price of Cyber (In)security: Evidence from the Italian Private Sector. *Bank of Italy Occasional Paper 407.* Rome: Bank of Italy.

Biancotti, Claudia, and Paolo Ciocca. 2018. Data Superpowers in the age of AI: A Research Agenda. *Voxeu.org* (October 23).

Boddington, Paula. 2017. *Towards a Code of Ethics for Artificial Intelligence.* Springer.

Bohannon, John. 2015. Privacy. Credit card study blows hole in anonymity. *Science* 347 (6221).

Bourreau, Marc, Alexandre de Streel, and Inge Graef. 2017. *Big Data and Competition Policy: Market Power, Personalized Pricing and Advertising*. Brussels: Centre on Regulation in Europe (CERRE).

Chirita, Anca. 2019 (forthcoming). Data-Driven Mergers Under Competition Law. In *The Future of Commercial Law: Ways Forward for Harmonisation*, ed. Orkun Akseli and John Linarelli. Oxford: Bloomsbury.

Cockburn, Iain M., Rebecca Henderson, and Scott Stern. 2018. The Impact of Artificial Intelligence on Innovation. In *The Economics of Artificial Intelligence: An Agenda*, ed. Ajay Agrawal, Avi Goldfarb, and Joshua Gans. Cambridge, MA: National Bureau of Economic Research.

Demchak, Chris, and Yuval Shavitt. 2018. China's Maxim—Leave No Access Point Unexploited: The Hidden Story of China Telecom's BGP Hijacking. *Military Cyber Affairs* 3, no. 1.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A Large-Scale Hierarchical Image Database. *2009 IEEE Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Executive Office of the President of the United States. 2014. *Big Data: Seizing Opportunities, Preserving Values*. Washington: The White House.

Garfinkel, Simson. 2015. *De-Identification of Personal Information*. National Institute of Standards and Technology (NIST) Interagency/Internal Report (NISTIR) 8053. Washington: National Institute of Standards and Technology, US Department of Commerce.

Graef, Inge. 2016. *EU Competition Law, Data Protection and Online Platforms. Data as Essential Facility*. Alphen aan den Rijn, Netherlands: Wolters Kluwer.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–78. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Huth, Alexander G., Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. 2016. Natural Speech Reveals the Semantic Maps that Tile Human Cerebral Cortex. *Nature* 532: 453–58.

Kokolakis, Spyros. 2017. Privacy Attitudes and Privacy Behaviour: A Review of Current Research on the Privacy Paradox Phenomenon. *Computer Security* 64, 122–34.

Lanier, Jaron, and E. Glen Weyl. 2018. A Blueprint for a Better Digital Society. *Harvard Business Review* (September).

Mayer, Nikolaus, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2018. What Makes Good Synthetic Training Data for Learning Disparity and Optical Flow Estimation? *International Journal of Computer Vision* 126, no. 9: 942–60.

OECD (Organization for Economic Cooperation and Development). 2015. *Data-Driven Innovation: Big Data for Growth and Well-Being.* Paris.

Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni. 2016. The Synthetic Data Vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Pearl, Judea. 2018. *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution.* UCLA Technical Report R-475. University of California, Los Angeles.

Pitofsky, Robert, Donna Patterson, and Jonathan Hooks. 2002. The Essential Facilities Doctrine Under United States Antitrust Law. *Antitrust Law Journal* 70: 443–62.

Posner, Eric, and E. Glen Weyl. 2018. *Radical Markets*. Princeton, NJ: Princeton University Press.

Rosenblatt, Frank. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* 65, no. 6: 386–408.

Rubinfeld, Daniel L., and Michal Gal. 2017. Access Barriers to Big Data. *Arizona Law Review* 59: 339–81.

Schaub, Florian, Rebecca Balebako, and Lorrie Faith Cranor. 2017. Designing Effective Privacy Notices and Controls. *IEEE Internet Computing* 21, no. 3.

Torra, Vicenç, and Guillermo Navarro-Arribas. 2016. Big Data Privacy and Anonymization. In *Privacy and Identity Management: Facing up to Next Steps*, ed. Anja Lehmann. International Federation for Information Processing—Advances in Information and Communication Technology book series, volume 498. Springer.

This publication has been subjected to a prepublication peer review intended to ensure analytical quality. The views expressed are those of the authors. This publication is part of the overall program of the Peterson Institute for International Economics, as endorsed by its Board of Directors, but it does not necessarily reflect the views of individual members of the Board or of the Institute's staff or management.

The Peterson Institute for International Economics is a private nonpartisan, nonprofit institution for rigorous, intellectually open, and indepth study and discussion of international economic policy. Its purpose is to identify and analyze important issues to make globalization beneficial and sustainable for the people of the United States and the world, and then to develop and communicate practical new approaches for dealing with them. Its work is funded by a highly diverse group of philanthropic foundations, private corporations, and interested individuals, as well as income on its capital fund. About 35 percent of the Institute's resources in its latest fiscal year were provided by contributors from outside the United States. A list of all financial supporters is posted at https://piie.com/sites/default/files/supporters.pdf.