Consumer Privacy and Value of Consumer Data

Mehmet Canayaz*, Ilja Kantorovitch†, Roxana Mihet‡

August 19, 2022

Abstract

We analyze how the adoption of the California Consumer Privacy Act (CCPA), which limits consumer personal data acquisition, processing, and trade, affects voice-AI firms. To derive theoretical predictions, we use a general equilibrium model where firms produce intermediate goods using labor and data in the form of intangible capital, which can be traded subject to a cost representing regulatory and technical challenges. Firms differ in their ability to collect data internally, driven by the size of their customer base and reliance on data. When the introduction of the CCPA increases the cost of trading data, sophisticated firms with small customer bases are hit the hardest. Such firms have a low ability to collect in-house data and high reliance on data and cannot adequately substitute the previously externally purchased data. We utilize novel and hand-collected data on voice-AI firms to provide empirical support for our theoretical predictions. We empirically show that sophisticated firms with voice-AI products experience lower returns on assets than their industry peers after the introduction of the CCPA, and firms with weak customer bases experience the strongest distortionary effects.

Keywords: Data Regulation, Consumer Data, Data Governance, Firm Dynamics, Data and Finance, Internet Regulation, Competition Policy, Antitrust, Market Power

JEL-Codes: D80, G30, G31, G38, L20, O30

*Smeal College of Business, Penn State. Contact: mcanayaz@psu.edu

†EPFL. Contact: ilja.kantorovitch@epfl.ch

‡Swiss Finance Institute at HEC Lausanne. Contact: roxana.mihet@unil.ch

This draft: August 19, 2022. First version: October 22, 2021. *Acknowledgements*: We thank our discussants, Doh-Shin Jeon, Aija Leiponen, Antoine Uettwiller, Joel Waldvogel, as well as the audience at the Next Generation of Antitrust, Data Privacy and Data Protection Scholars Conference (NYU), 2022 NBER Economics of Privacy, 11th Biannual Toulouse Postal Economics Conference, 4th Future of Financial Information Conference, 2022 Young Swiss Economist Meeting, and 2022 SSES Annual Congress for their useful feedback. We also thank seminar participants at HEC Lausanne, ESSEC, and Penn State for their comments, as well as Jon Frost, Avi Goldfarb, Marc Painter, Christian Peukert, Thomas Philippon, and Laura Veldkamp for their constructive suggestions. Lastly, we thank Simona Abis and Huan Tang for their contribution to an earlier version of this paper. All errors are ours.

1 Introduction

Today's firms gather vast amounts of consumer data to design better and more innovative products, predict customer demand more accurately, increase operational efficiency, and implement well-targeted and profitable marketing campaigns.[1] The use of consumer data, however, also raises important challenges. One first-order concern is consumer privacy. In general, too strict consumer privacy laws may hinder firms, whereas full transparency of consumer data disregards privacy with possible discriminatory outcomes (Acquisti et al., 2016). Consumer privacy laws can have unintended consequences, as firms with previously-collected consumer data can be advantaged if the collection of consumer data or its purchase from third parties is restricted.

To understand the role of data restrictions on firm outcomes, we build a structural model in which firms accumulate data to increase the production of an intermediate good. We study multiple factors that can affect firm dynamics, such as firm-level differences in the ability to generate internal data and leverage data for production purposes (i.e., differences in the quality of the algorithms) and public policies toward data sharing. This parsimonious model allows us to obtain theoretical predictions on the possible effects of privacy regulations on firm and industry dynamics.

In our model, firms produce intermediate goods by combining labor and data. Firms can collect data internally or acquire data externally from other firms. We focus on several key frictions and dimensions of heterogeneity: First, firms differ in the amount of raw data they have about their customers, as some firms have larger customer bases or business models that amplify data collection. Second, firms can trade data subject to an iceberg transaction cost that changes according to regulations on data sharing between firms. Third, firms are heterogeneous in their data sophistication as expressed through their factor shares. We assume that some firms can leverage large amounts of data for production but are also more reliant on abundant data sources.

We use our model to study the effects of two ways through which the introduction of CCPA affects the ability of firms to use data. When CCPA impedes data sharing among firms through

[1] According to the International Data Corporation (IDC), global spending on big data and analytics solutions will increase to over 215 billion US Dollars in 2021, a 10% increase from 2020. The IDC forecasts a compound annual growth rate of 12.8% between 2021 and 2025 (Vesset and George, 2021). This reflects an increase in the volume of data globally generated data, which increased exponentially and reached 33 Zettabytes in 2018 with a projected volume of 175 Zettabytes in 2025 (Rydning, 2018).

an increase in the iceberg transaction cost, we find that all firms are hurt, but sophisticated firms with small customer bases are hurt the most. In contrast, firms with large customer bases are the least affected. The reason is that the cost of trading frictions is paid mainly by data buyers, as their demand for data is inelastic relative to supply. This is because data buyers cannot adequately substitute internal data for external data and are generally more reliant on large quantities of data. Moreover, the non-rivalry of data makes the data supply more elastic. In contrast, privacy policies that limit the ability of firms to collect data on their customers, sophisticated firms, both large and small, are similarly negatively affected, as data becomes overall scarcer. These results indicate that public policies that limit firms' data sharing distort competition between small firms and large incumbents more than policies affecting data accumulation because the former hurt small sophisticated firms.

We test this theory and add to the literature by exploiting a novel and hand-collected data set of 15,642 conversational-AI firms with access to detailed voice-generated data on U.S. consumers between January 2017 and February 2022. These firms operate through 27,614 unique voice assistant products on electronic devices. They listen to customers when triggered, exploit recorded data about them, and receive instantaneous feedback through speech recognition and natural language understanding technologies. Since these firms communicate with users vocally, they can build detailed customer profiles and infer personal habits, gender, and ethnic background.

We analyze how adopting the California Consumer Privacy Act (CCPA) impacted these previously unstudied firms. According to the State of California Attorney General's Office, the CCPA "gives consumers control over the personal information businesses collect on them and how it is used and shared." Moreover, it gives consumers the right to choose not to have their data re-sold to third parties, the right to demand businesses erase their personal information, and the right to be not discriminated against for exercising any of the above.

In a first step, our empirical approach draws inferences based on comparing firms with voice-AI products to their industry peers before and after the introduction of the CCPA. In a second step, we differentiate within this group of sophisticated firms and focus on firms with and without in-house data. We define firms with in-house data as firms that have gathered more customer feedback per voice-AI product than their competitors and proxy for such feedback through average ratings and

the number of reviews. Our panel contains 30,134 observations at the firm-quarter level. This allows us to introduce firm and industry x quarter fixed effects that control for fixed characteristics at the firm and industry-quarter levels, which may otherwise be confounders to the relationship between the CCPA and product outcomes.

We find that the adoption of the CCPA affects sophisticated firms with voice-AI products negatively relative to firms without voice-AI products through lower returns on assets. Such firms experience a 1.59% lower return on assets (t-stat = -2.36). This negative effect is even larger for firms without in-house data as proxied by low ratings or few reviews. Such sophisticated firms with small customer bases experience a decline in return on assets of up to -2.87% (t-stat = -3.74). These results confirm the theoretical predictions that restrictions on data trade hit sophisticated firms with small customer bases the hardest. Moreover, these results are robust to controlling for debt to assets, log assets, and firm age.

One concern related to the estimation of CCPA's effects on voice-AI firms is the potential violation of the parallel trends assumption, for example, due to anticipation or lobbying. Although we provide an abundance of evidence on the observed counterparts of the parallel trends assumption (see Figure 8 among others), we still complement our empirical findings by running double machine learning (Double ML) techniques that model treatment and estimate heterogeneous conditional treatment effects following Robinson (1988) and Chernozhukov et al. (2016, 2017). Our results from running Double ML regressions provide additional evidence for the distortionary effects of the CCPA on firms with less in-house data, along with young and small firms.

This paper pushes the knowledge frontier in several ways. First, we gather a unique data set on firms that have access to valuable personal information on U.S. consumers in the form of voice-generated data. Second, we exploit the introduction of the CCPA as it is a data privacy regulation that accounts for best data practices within the newest technologies.[2] While studies have been conducted to explore business ramifications of privacy regulations, no paper has attempted to determine the impact, in terms of firm profitability, of the CCPA legislation itself. Third, we

[2]Older data privacy laws such as the Family Education Rights and Privacy Act (FERPA) and the Health Information Portability and Accountability Act (HIPAA), studied by Khansa et al. (2012) among others, are still important and useful, but they do not account for newer information technologies or for the different ways that data is being used to predict consumer preferences or to price-discriminate.

identify the channels through which firms' financial decisions and outcomes change in response to data restrictions using our novel data set of conversational-AI firms and the CCPA as a shock to both customer acquisition and to data processing ability. Lastly, we show that our empirical findings confirm the theoretical predictions of a model in which firms differ in their ability to produce data internally combined with trading frictions.

2 Literature Review

We contribute to the literature that studies the impact of the newest data regulations on firm performance. It has been reported that the GDPR hurt the performance of firms in the airline industry (Aridor et al., 2020), and on European firms' ability to attract investment (Jia et al., 2021). Moreover, the GDPR seems to have negatively impacted innovation since AI startups are re-allocating their limited resources to deal with the implications of the GDPR (Bessen et al., 2020). However, some studies find a neutral or even positive effect of the GDPR. Godinho de Matos and Adjerid (2021) show that consumers' opt-in decisions for a large European telecommunications provider have increased after the GDPR, leading to increased sales due to more effective targeted advertising. We find a similar effect of higher profitability and higher customer ratings in the aftermath of CCPA, but only for firms with more in-house data.

Our analysis contributes to the debate on whether the use of new information technologies such as AI and big data distorts competition dynamics and ends up creating winner-takes-all effects. Data can empower growth and innovation because data has the particular economic property of non-rivalry (Jones and Tonetti, 2020; Cong et al., 2021), which means that data can be used by any number of firms simultaneously without losing its value, e.g., for the training of algorithms. However, precisely this economic property of data can lead to anti-competitive effects. For example, (Farboodi et al., 2019) has warned about the emergence of data-feedback loops that allow large firms to grow even larger. Large firms generate more data as a byproduct of economic activity. When combined with technical sophistication and good algorithms, larger firms succeed at making better predictions, producing more efficiently, and expanding faster than smaller firms. Hagiu and Wright (2020) argues that when the marginal value of learning from customer data remains high even after

a very large customer base has been acquired, this can lead to large and persistent competitive advantages. Data may also allow large firms to expand into new markets (Vives and Ye, 2021) more easily.

Indeed, some of these theoretical effects have been verified empirically as well. For example, Babina et al. (2021) find that firms that invest more in AI experience faster growth in sales and employment, and the AI growth effects concentrate among the ex-ante largest firms, leading to higher industry concentration and reinforcing winner-take-most dynamics. Moreover, Hoberg and Phillips (2021) find that over the past years, U.S. firms have expanded their scope of operations. Increases in scope and scale were achieved largely without increasing traditional operating segments but rather through scope expansion, primarily realized through acquisitions and investment in R&D and not through capital expenditures.

Data regulations could end these winner-take-most effects, ameliorate market competition by encouraging firm entry (Babina et al., 2022) and avoid other harmful societal effects of unregulated AI (Acemoglu, 2021; Acquisti et al., 2016). Our paper shows that data restrictions, such as the CCPA, hurt firms without in-house data the most, as they restrict the ability of firms to buy external data (for a model of data intermediation, see Bergemann et al., 2020). Firms with in-house data already have deep knowledge about their typical customers, so data restrictions do not hurt the performance of their AI algorithms as much. As a result, firms with in-house data expand their market share. This is in line with the theoretical findings of Eeckhout and Veldkamp (2021) who warn that big firms are using data to reallocate production to the goods consumers want most. As large firms already have this data, it is easier for them, relative to young, small firms, to tailor their products to consumers' preferences. Our work parallels the contemporaneous findings of Babina et al. (2022) who argue that unequal data access gives incumbents an advantage by discouraging new entry, as well as the work of Peukert et al. (2021) and Campbell et al. (2015), who show that data regulation can create barriers to entry and may thus hurt competition. In particular, Campbell et al. (2015) show that though privacy regulation imposes costs on all firms, it is small firms and new firms that are most adversely affected, particularly for goods where the price mechanism does not mediate the effect, such as the advertising-supported internet. Similarly, Peukert et al. (2021) find that while all firms suffer losses from the GDPR, the dominant vendor, Google, loses relatively less

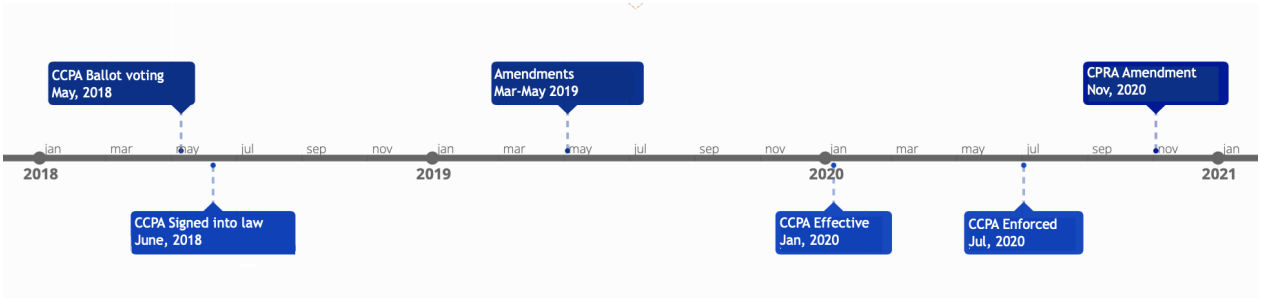
and can significantly increase market share in important markets such as advertising and analytics.

The remainder of the paper proceeds as follows. In section 3 the institutional details of the CCPA are presented and put in a context to Voice-AI products and customer privacy. Section 4 then presents a theory that motivates our empirical approach. Then section 6 presents our data and section 7 lays out our empirical strategy and findings. Finally, section 9 concludes.

3 The California Consumer Privacy Act

According to the State of California Office of the Attorney General (https://oag.ca.gov), the California Consumer Privacy Act (CCPA) went through a fast timeline of only 20 months from the first ballot voting in May 2018 to its enforcement date in July 2020, as shown in Figure 1. Various significant amendments have also come into effect during this process. An amendment made in April 2020 exempted "insurance institutions, agents, and support organizations" from the law because they were already subject to similar regulations under California's Insurance Information and Privacy Protection Act (IIPPA). Another significant amendment, the California Privacy Rights Act (CPRA), often called CCPA 2.0, was approved in November 2020. The CPRA adds new provisions related to establishing a California Privacy Protection Agency.

Figure 1: **CCPA Timeline:** From Signing into Law to Effective Data
Source: Office of the Attorney General.



We use the implementation of the CCPA for our analysis because it dramatically alters the way U.S.-based companies process data. The law includes detailed disclosure requirements, provides individuals with extensive rights to control how their personal information is used, imposes statutory fines, and creates a private right of action. CCPA defines "personal information" much more broadly

than it is defined under most U.S. privacy laws. It is defined as any information that could reasonably be linked to a particular person or household, whether directly or indirectly. This includes real name, physical address, biometric information, IP address, online identifier, license number, passport number, race, records of purchasing history or tendencies, internet browsing and search history, geolocation data, audio data, employment, or education data, as well as inferences drawn from these.

Different from centralized opt-outs, e.g., iOS's "Do Not Track",[3] tracking the share of consumers exercising their CCPA rights is more complicated and relies on firms disclosing this information. According to the IAB, a New York-based consortium of advertising, publishing, and marketing enterprises surveyed privacy attorneys in the industry,[4] which states that the CCPA has effects far beyond California: 60% of responding firms state that they provide CCPA rights uniformly, i.e., also to customers in the US outside of California or customers in the EU. This broad effect of the CCPA is contrasted by a low percentage (1% - 5%) of consumers exercising their right to opt-out of data selling across channels. Moreover, numerous firms have stopped their data selling activities in anticipation of CCPA. DataGrail, a data privacy startup, confirms that privacy rights are exercised outside of California, highlighting that companies tend to offer CCPA privacy rights for all US consumers.[5] Moreover, DataGrail tracked on average 266 privacy requests per one million identities in 2021, of which 63% were requests not to sell data.

3.1 Conversational-AI and Customer Privacy

Amazon is the largest online retailer in the U.S., generating more than $457 billion in sales in 2021.[6] Founded in July 1994, the company first launched its digital assistant, the Alexa smart

[3]According to Flurry, an app analytics company, the vast majority of users have not opted-in to tracking across apps months after the launch of iOS 14.5, which introduced the global switch.

[4]See the IAB CCPA Benchmark Survey: https://www.iab.com/wp-content/uploads/2020/11/IAB_CCPA_Benchmark_Survey_Summary_2020-11.pdf.

[5]See https://www.datagrail.io/resources/reports/2022-ccpa-trends-report/ for DataGrail's report on CCPA.

[6]See, e.g., macrotrends.net. In terms of sales, Amazon leads in global smart speaker sales and continues to expand its lead over Google and Apple, according to VoiceBot.AI. Amazon sold 16.5 million smart speakers and smart displays during the period 2019-2020, followed by Google with 13.2 million, Baidu at 6.6 million, and Alibaba with 6.3 million. Apple came in the fifth slot with 4.6 million smart speakers sold in the fourth quarter.

speaker, in 2014, which was made available to the general public in 2015, three years before CCPA was announced and five years before CCPA was passed. Amazon's Alexa dominates the market for voice-AI around the globe. In 2018, Amazon Alexa had a 72% market share, compared to 18.4% by Google in the U.S., where the market for smart speakers has been growing at a 30-40% annual rate. An analysis from Voicebot.ai shows that the number of adults using smart-speakers grew from 47 million in 2018 to 90 million in 2020, which is approximately 35% of the U.S. population.

Alexa's Skills, i.e., voice-AI products, allow customers to use their voices to perform everyday tasks such as checking the news, listening to music, playing games, shopping, accessing news services, scheduling transactions, checking their bank balances, performing financial transactions, or controlling other smart home devices and other utilities. Companies and individuals can publish Skills in the Alexa Skills Store to reach users of Alexa devices. Amazon made Alexa's Application Programming Interface (API) openly available to developers, allowing for integration in non-Amazon devices. Developers can interact with Alexa by developing Alexa Skills, integrating Alexa with third-party hardware, or adding Alexa support to IoT hardware. Businesses can use Alexa's existing capabilities to accomplish business tasks or build new Skills and integrations through the "Alexa for Business" platform. In short, the Alexa platform allows businesses to "always listen" to customers from Amazon devices or third-party tablets or phones and collect extremely valuable information about them. This extensive data collection is often used for improving products for the customers but can also constitute a violation of their privacy.

Although most Alexa Skills are free, there are many ways for firms to earn money by creating an Alexa Skill.[7] First, Alexa Skills can be used to order goods and services directly through the skill.[8] Examples are food delivery or ride-hailing services. Second, Alexa Skills can offer in-skill purchases to unlock premium features, for example, additional clues in a game or interactive stories. Finally, firms may want to provide Alexa Skills to strengthen their brand and collect data on the preferences and needs of customers. Users can link their third-party accounts with Alexa Skills to access a broader set of features.[9] Account linking allows firms to gather additional customer information and further complement existing customer profiles.

[7] See the Amazon Alexa Skill manual.
[8] See the overview for direct purchases of goods and services through Alexa Skills.
[9] See the account-linking manual.

4 Model

In this section, we build a simple theoretical model to investigate the effects of privacy regulations on firm outcomes, focusing mainly on their data accumulation process. We present a monopolistically-competitive economy in which firms accumulate data to increase the production of an intermediate good. The novel and crucial feature of the model is that data accumulation results from internal data generation as a byproduct of economic activity and external data acquisition from third parties through data trade.

We separate the possible effects of privacy regulations on firms' economic outcomes by considering a multitude of shocks to firms' data accumulation process. We explore shocks to the cost of external data acquisition as privacy laws prohibit firms from sharing user data with third parties without active consent, thus increasing the direct and indirect costs of external data trading between firms.[10] We also explore shocks to firms' ability to gather data about their own consumers, because privacy protection laws limit the amount of personal data firms can collect on their customers, as data collection requires valid affirmative consent. Valid consent makes data collection more expensive for firms and can thus reduce a firm's ability to acquire customers and gather internal data as a byproduct of sales.

Households There is a unit mass of households with utility

$$U = \ln(C). \tag{1}$$

Each household consists of a worker and a data analyst who supply their unit of labor in-elastically, earn wages and consume. The household side of the model is kept simple as the main focus lies on the firms.

Final Good Sector The final good and numéraire is assembled by many firms using intermediate goods with a CES technology

$$Y = \left(\int_0^1 Y_i^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}}, \tag{2}$$

[10]The primary contribution of this theoretical section is to provide a simple framework to think about the firm value of consumer data and the consequences that data accumulation and privacy laws have on firm and industry dynamics.

where $\sigma > 1$ is the elasticity of substitution. The price for each intermediate good is determined competitively and equal to the marginal product of intermediate good i,

$$P_i = \frac{\partial Y}{\partial Y_i} = \left(\frac{Y}{Y_i}\right)^{\frac{1}{\sigma}}. \tag{3}$$

Intermediate Good Sector Firms produce intermediate goods by combining labor with data through a Cobb-Douglas production function

$$Y_i = D_i^{\alpha_i} l_i^{1-\alpha_i}, \tag{4}$$

where

$$D_i = \left(\left(D_i^I\right)^{\frac{\varepsilon-1}{\varepsilon}} + \xi\left(D_i^E\right)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{\varepsilon-1}} \tag{5}$$

is a data bundle combining internal data D_i^I and external data D_i^E, l_i is labor from workers, and α_i is the factor share of the data bundle. The elasticity of substitution between internal and external data is $\varepsilon > 1$. Generally, external data may be less productive than internal data incorporated through $\xi \leq 1$. Data makes labor more productive, as goods are tailored to the customer's needs, and production techniques are improved.

Internal data is produced within the firm,

$$D_i^G = A_i^G (l_i^G)^{1-\phi}, \tag{6}$$

where A_i^G stands for the size of a firm's customer base, which facilitates data collection, l_i^G is labor from data analysts. We can think that data analysts need to implement routines to capture data from user actions and make it usable, such that internal data is increasing in the amount of labor dedicated to data gathering.

Firms can decide to share their data with other firms. Such data trade has two crucial features in this model.

1. Non-rivalry: Firms keep a fraction $\nu \in [0, 1]$ of shared data.

2. Trade frictions: For any unit of data shared, only $1-\tau \in [0, 1]$ units arrive (*iceberg transaction*

costs).

First, partial non-rivalry is a real feature of data. Firms do not lose access to shared data. Still, shared data may be less useful as having access to unique data may increase the firm's market power. As a result, total internal data is equal to

$$D_i^I = D_i^G - (1-\nu)D_i^S, \tag{7}$$

where D_i^S denotes data shared with other firms, and ν denotes the degree of non-rivalry in data trade. We capture this loss of the value of data through sharing through a decrease in the quantity of internal data.

Second, trade frictions are central to our analysis and stand in for restrictions to data sharing that stem from *privacy laws*. Privacy laws may create additional regulatory hurdles to the trade of data, which we capture through the proportional cost τ.

On a more conceptual note, we think of data sharing as firm i deciding with how many firms to share its data. As data is shared with more and more firms, the data increasingly loses value for all firms involved. As D_i^S approaches $\frac{D_i^G}{1-\nu}$, the value of the shared data vanishes for the individual firm, but the total market value of data sharing for a firm i is maximized. We do not consider data sharing that exceeds this point and also reduces the market value of data sharing for a firm i.

Taking all together, the firm's maximization problem is

$$\max_{l_i, l_i^G, D_i^S, D_i^E} \quad Y^{\frac{1}{\sigma}} D_i^{\alpha_{D,i}} l_i^{\alpha_{L,i}} - w_Y l_i - w_G l_i^G + p^D D_i^S - \frac{p^D}{1-\tau} D_i^E \tag{8}$$

$$s.t. \quad D_i = \left(\left(D_i^I\right)^{\frac{\varepsilon-1}{\varepsilon}} + \xi \left(D_i^E\right)^{\frac{\varepsilon-1}{\varepsilon}} \right)^{\frac{\varepsilon}{\varepsilon-1}} \tag{9}$$

$$D_i^I = D_i^G - (1-\nu)D_i^S \tag{10}$$

$$D_i^G = A_i^G \left(l_i^G\right)^{1-\phi} \tag{11}$$

$$l_i, l_i^G, D_i^E, D_i^I, D_i^S \geq 0 \tag{12}$$

where firm revenue $\Pi_i = P_i Y_i = Y^{\frac{1}{\sigma}} D_i^{\alpha_{D,i}} l_i^{\alpha_{L,i}}$ has already been written out, w_Y is the wage rate for workers and w_G is the wage rate for data analysts. The trade friction τ drives a wedge between the

price of selling one unit of data p^D and buying one unit $\frac{p^D}{1-\tau}$. Effective factor shares are $\alpha_{D,i} = \frac{\sigma-1}{\sigma}\alpha_i$ and $\alpha_{L,i} = \frac{\sigma-1}{\sigma}(1-\alpha_i)$.

We consider two dimensions of heterogeneity among firms. First, as in Farboodi et al. (2019), firms differ in their ability to generate data internally, i.e., have different A_i^G. A high A_i^G may stand for a *large customer base* or *a business model that amplifies data collection.* Moreover, we consider firms that differ in their *sophistication*, i.e., ability to process large amounts of data, but also reliance on data for their operations. Sophisticated firms have a larger $\alpha_{D,i}$. In connection to our empirical analysis, we identify firms with in-house data as firms that have larger customer bases or high ability to generate data internally through a large A_i^G, whereas differences in α_i suggest which type of firms may suffer more adverse effects from data restrictions. Both dimensions of heterogeneity can be combined, such that unsophisticated firms may have large customer bases, and firms with small customer bases may be sophisticated.

The first order conditions are given by,

$$(l_i) : \alpha_{L,i}\frac{\Pi_i}{l_i} = w_Y \tag{13}$$

$$\left(l_i^G\right) : \alpha_{D,i}\left(1-\phi\right)\frac{\Pi_i}{D_i}\left(\frac{D_i}{D_i^I}\right)^{\frac{1}{\varepsilon}}\frac{D_i^G}{l_i^G} = w_G \tag{14}$$

$$\left(D_i^S\right) : \alpha_{D,i}\left(1-\nu\right)\frac{\Pi_i}{D_i}\left(\frac{D_i}{D_i^I}\right)^{\frac{1}{\varepsilon}} = p^D - \lambda_i \tag{15}$$

$$\left(D_i^E\right) : \alpha_{D,i}\xi\frac{\Pi_i}{D_i}\left(\frac{D_i}{D_i^E}\right)^{\frac{1}{\varepsilon}} = \frac{p^D}{1-\tau} \tag{16}$$

where λ_i is a Lagrange multiplier for the non-negativity condition of data sharing D_i^S with the corresponding slackness condition

$$\lambda_i D_i^S = 0. \tag{17}$$

Whenever firm i finds it optimal to share some data ($D_i^S > 0$), λ_i equals zero, and $\lambda_i > 0$ whenever firm i decides not to share any data. In general, firms that have a larger customer base A_i^G will share data, whereas firms with small customer bases will not share data if ν is sufficiently small.

Market Clearing Before moving to our comparative statics, we close the model through the

market clearing conditions.

$$\int_0^1 l_i di = 1 \tag{18}$$

$$\int_0^1 l_i^G di = 1. \tag{19}$$

$$\int_0^1 D_i^S di = \int_0^1 D_i^E di + \tau \int_0^1 D_i^S di, \tag{20}$$

$$Y = w_Y \int_0^1 l_i di + w_G \int_0^1 l_i^G di + \int_0^1 \pi_i di. \tag{21}$$

The first two conditions pin down wages w_Y and w_G for workers and data analysts. The third condition states that all shared data equals external data plus the iceberg transaction costs, and the resulting price of data p^D clears the market. The final condition states that households use their wages and firm profits $\int \pi_i di$ to buy all final goods.

5 Model Results

We focus on two effects of privacy regulations. Firms may need to ask for explicit approval to engage in data sharing. Staying compliant with restrictions on data trade leads to trade frictions as captured by an increase in τ. Second, gathering internal data may also become more complicated, corresponding to a decrease in A^G. Such regulation could have heterogeneous effects on firms depending on their ability to gather internal data (A_i^G) and their dependence on data ($\alpha_{D,i}$).

5.1 Theoretical Findings

Before moving to the numerical simulations, consider the simplified setting where firms with large customer bases do not value external data ($\xi = 0$). Firms with small customer bases do not share data and treat internal and external data as perfect substitutes ($\varepsilon \to \infty$). Denote Π_i^S as the profit of data sharing firms and Π_i^B as the profit of data buying firms. Using the envelope theorem, comparative statics of profits for the trading friction τ are given by

$$\frac{\partial \Pi_i^S}{\partial \tau} = \frac{\partial Y^{\frac{1}{\sigma}}}{\partial \tau} D_i^{\alpha_D} l_i^{\alpha_L} + \frac{\partial p^D}{\partial \tau} D_i^S \tag{22}$$

$$\frac{\partial \Pi_i^B}{\partial \tau} = \underbrace{\frac{\partial Y^{\frac{1}{\sigma}}}{\partial \tau} D_i^{\alpha_D} l_i^{\alpha_L}}_{\text{GE effect}} - \underbrace{\frac{\partial \frac{p^D}{1-\tau}}{\partial \tau} D_i^E}_{\text{Data Trade}}. \tag{23}$$

There are two effects on firm profits. The general equilibrium effects work through aggregate demand Y. As a reduction in data trade keeps data from being employed where it is most productive, it must be that Y is decreasing in trading frictions τ. This effect is negative for all firms and depends on firm size. The data trade effect works through the effect on the price of data selling p^D and data buying $\frac{p^D}{1-\tau}$. An increase in $\frac{p^D}{1-\tau}$ leads to a fall in demand for data by buying firms. As the supply is initially unchanged, p^D needs to adjust downward to clear the market. In equilibrium, $\frac{p^D}{1-\tau}$ must increase to sustain the fall in demand.

Whether the price of buying or selling data adjusts more strongly depends on the elasticity of supply of and demand for data. When supply is perfectly elastic, p^D is pinned down, and $\frac{p^D}{1-\tau}$ has to adjust, putting all costs on data buying firms. In contrast, when demand is very elastic, small increases in $\frac{p^D}{1-\tau}$ can lead to a collapse in demand, leading to a substantial fall in p^D.

In our setting, data-selling firms are larger and can gather data at a low cost, leading to a high elasticity of the data supply. Partial non-rivalry makes the data supply even more elastic as the cost of data sharing shrinks. Indeed, with perfect non-rivalry, non-strategic firms share as much data as possible independent of the price. In contrast, firms with small customer bases cannot easily substitute for externally-acquired data, as their ability to generate data in-house is severely limited. This leads to inelastic demand for data. As a result, p^D adjusts by less than $\frac{p^D}{1-\tau}$. When the data supply is sufficiently elastic relative to demand, firms with small customer bases may experience a larger fall in profits than firms with large customer bases. Note that sophisticated data buyers rely more on data through a larger $\alpha_{D,i}$ and, therefore, acquire more data externally. As a result, sophisticated small firms experience a larger fall in profits than unsophisticated firms. The general equilibrium effect leads to a fall for all profits proportional to firm size.

5.2 Numerical Simulation

In our full model, firms are free to simultaneously buy and sell data, as in-house data can be helpful to learn about one's customers, and externally-acquired data can give information about new

potential customers or add information. We find that for the considered calibration, the theoretical findings carry over to the more general setting when considering relative changes in profits, which is a more appropriate metric when firms differ in size. First, we see that the price of data buying $\frac{p^D}{1-\tau}$ reacts stronger than p^D, indicating that data supply is more elastic than demand. As a result, small firms suffer more from trading frictions as they rely more on external data. Moreover, sophisticated firms of all sizes suffer more than unsophisticated firms as they buy more and share less data.

Calibration In the following, we focus on a calibration where the parameters for the simulations are given by $\alpha_{unsoph} = 0.2$, $\alpha_{soph} = 0.4$, $\epsilon = 2$, $\xi = 1.0$, $\nu = 0.1$, $\phi = 0.3$, and $A^G_{small} = 0.2 * A^G$ and $A^G_{large} = 1.0 * A^G$. Moreover, we assume that half of the firms have large customer bases, whereas half have small ones. Each group is also evenly split between sophisticated and unsophisticated. When we study changes in trading friction τ, we fix the common factor A^G to one. When we study changes to the firms' ability to generate data internally, we vary the common factor A^G and fix $\tau = 0.2$. The non-rivalry parameter is fixed to $\nu = 0.1$ throughout.

5.2.1 Impact of data sharing frictions

As a first exercise, we study the effect of restrictions on data trade, i.e., an increase in τ. The direct effect is that it becomes more costly to access external data, and firms substitute missing external data by increasing internal data gathering. However, not all firms are in the same position to increase their internal data gathering. In particular, firms with a larger customer base A^G_i can easily increase their data gathering. Moreover, the effect should be amplified for sophisticated firms that rely more on data. The result on profits is captured in Figure 2.

When looking directly at how firms adjust their data collection in Figure 3, we see that all firms reduce the acquisition of external data and increase their internal data collection, confirming the channel through which we see the dispersion in profits.

In general equilibrium, increasing trading frictions τ make the economy less productive, as data cannot flow to where it is most needed. As a result, the price of data p^D, wages w_Y and w_G, and output Y fall. This result is captured in Figure 4. In contrast, the price of data buying $\frac{p^D}{1-\tau}$ increases.

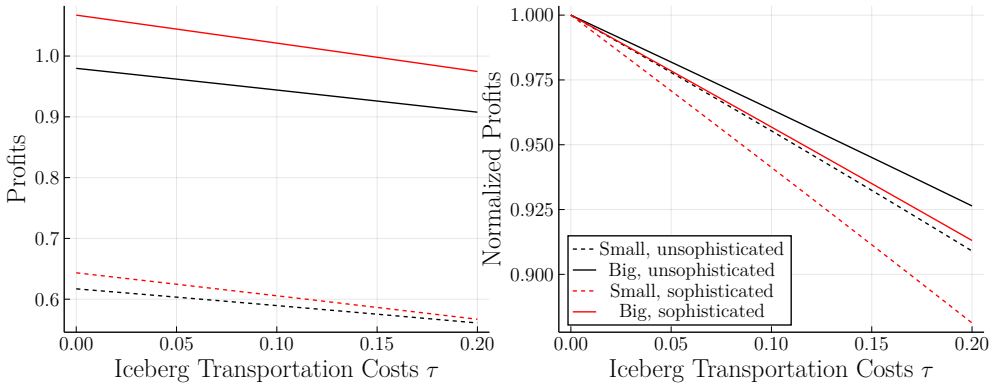


Figure 2: **Firm Profits and Higher Iceberg Transaction Costs:** All firms are negatively affected by worse access to external data, but sophisticated firms with small customer bases are hit the hardest.

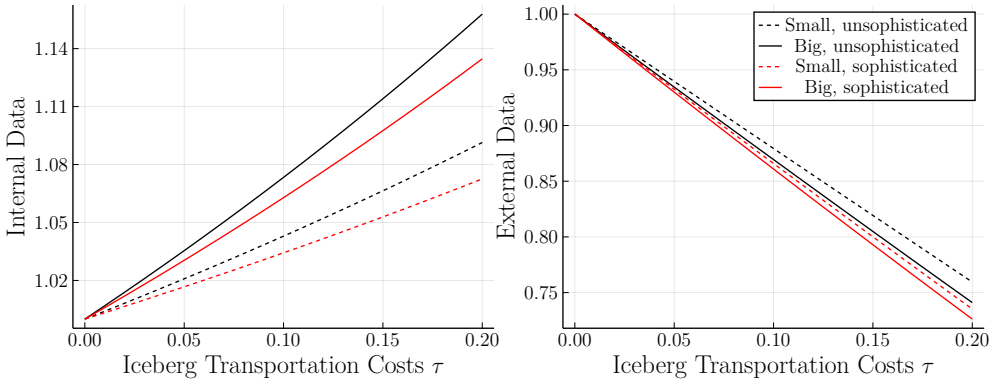


Figure 3: **Relative Firm Responses to Higher Iceberg Transaction Costs:** Firms try to substitute missing external data by generating more internal data, which is easier for firms with large customer bases.

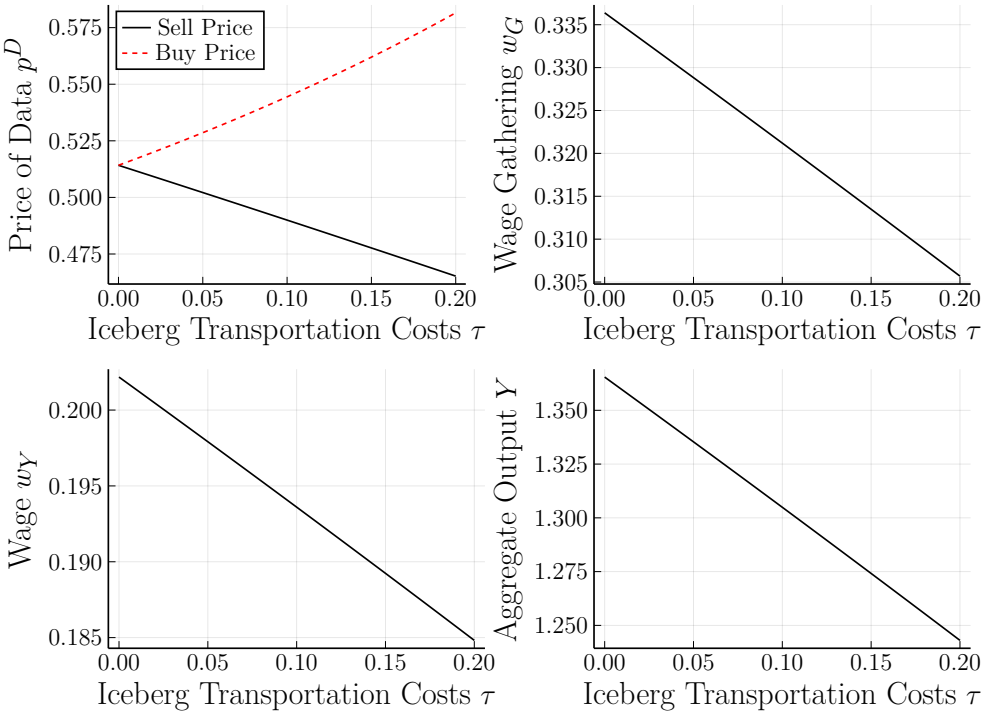


Figure 4: **Aggregate Responses to Higher** τ. An increase in the iceberg transaction cost τ leads to a fall in the price of data p^D, in wages w_Y and w_G and in lower aggregate output Y.

5.2.2 Impact of lower ability to generate in-house data

Whereas changes in the trading friction τ affected the trade in data, favoring internal data generation compared to the acquisition of external data, changes in the ability to generate internal data A_G mainly make data scarcer overall. Scarcer data hurts all firms but sophisticated firms more severely, which rely more on data. However, data can still flow from firms with large customer bases to smaller firms. Therefore, the main dimension of heterogeneity is sophistication rather than size. This result is captured in Figure 5.

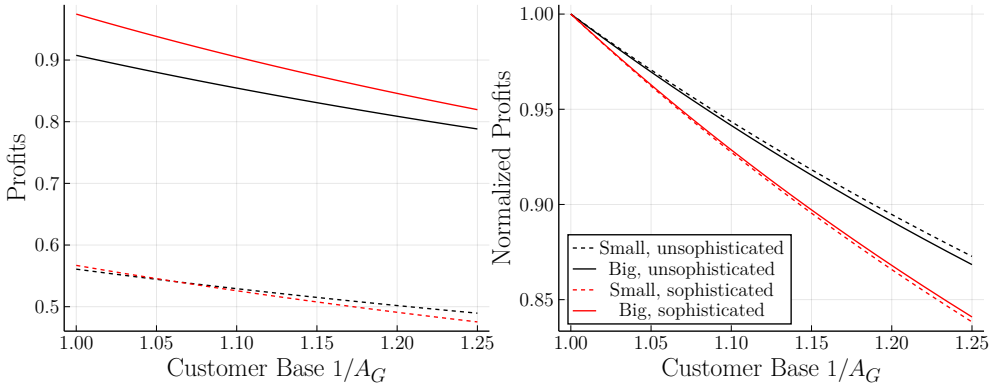


Figure 5: **Firm Profits and Data Generation:** The profits of all firms fall as data becomes scarcer. As sophisticated firms rely more on data, their profits fall the most.

Indeed, firms with small or large customer bases reduce their acquisition of external data by similar amounts in relative terms, avoiding the additional negative effect on small firms. This result is captured in Figure 6.

The general equilibrium effect of restricted data gathering is that data becomes scarcer, which increases the price of data p^D. The economy also becomes less productive, decreasing wages and output. This result is captured in Figure 7.

5.3 Model take-aways

To summarize, our theoretical framework predicts that a tightening of data regulations hurts *all* firms, but to *different* degrees. The technologically-sophisticated firms with small customer bases suffer the most when data trading is affected. In that sense, data regulations may entrench incumbents because their small but savvy competitors, who could otherwise grow to compete against

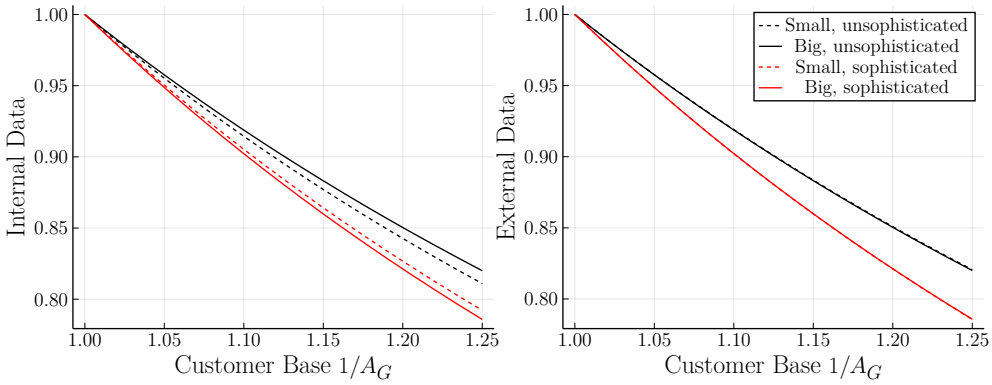


Figure 6: **Relative Firm Responses to Lower Data Generation:** Heterogeneous effects on firms are driven by the difference between sophisticated and unsophisticated firms, whereas firms with the same level of sophistication but different customer bases are similarly affected.

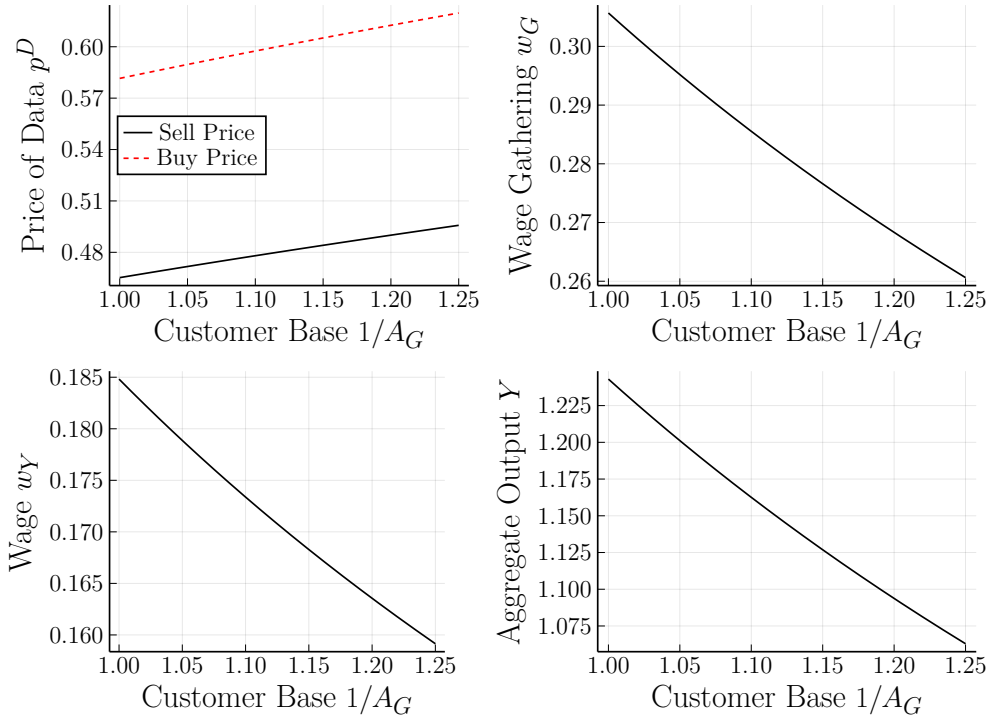


Figure 7: **Aggregate Responses to Lower** A_G. Less data generation increases the price of data and makes labor less productive.

the large incumbents, stand to lose the most from data restrictions. The main reason is that *sophisticated* firms are better at leveraging large amounts of data but need data, to begin with, to grow. Not having internal data or being unable to substitute internal data with external data adequately hurts them the most. The heterogeneous effect on small firms is not present when only data generation is affected. In that case, sophisticated firms are hit the hardest, with the size of the customer base only playing a minor role.

6 Data

This section presents our data. We build a unique data set on businesses that collect vast amounts of consumer personal information by scraping every product in the United States that has ever been listed on the Amazon Alexa Skills website. For each product, we have information on its date of entry into the conversational-AI space, its consumer reviews expressed on a scale from 1 (poor) to 5 (good), and its textual reviews. We merge this unstructured data set with financial and accounting information from CRSP/Compustat.

6.1 Descriptive Statistics

We construct a panel of firms and voice-AI products (Alexa Skills) by scraping all of Amazon's Alexa universe between January 2017 and February 2022. Our data on Alexa Skills is daily and contains 26,443,950 observations. For each Skill, we have a unique product identifier called **A**mazon **S**tandard **I**dentification **N**umber (ASIN), the name of its manufacturer, a customer rating between zero and five, and the number of verified customer reviews. The Alexa data is merged with the Compustat universe professionally by Effixis.[11] Our U.S. sample contains 381 unique gvkeys. For the merged firms, we utilize data from WRDS Financial Ratio Suite. The code to generate the variables of interest is available for download at https://whr.tn/3AtgfhC/. We keep firms with non-missing observations between 2017Q1 and 2020Q3. To develop a better control group, we drop all firms from the Fama-French-30 industries that are not represented in our Alexa sample along with firms from the Utilities and Trading sectors. We also drop all Alexa firms that do not have

[11] See https://effixis.ch/. Additional details on our data merge are available upon request.

active Alexa Skills as of 2017Q4, i.e., two quarters before the introduction and three years before the adoption of the CCPA. These restrictions yield a final panel of 1,586 unique gvkeys and 19 quarters, and therefore 30,134 observations in total.

Table 1: **Summary Statistics**

This table reports summary statistics on voice-AI firms. Financial data is from WRDS Financial Ratio Suite. D2AT denotes debt to asset ratio, and ROA denotes return on assets in percentage terms. Log(AT) and Log(AT)^2 are log book value of assets and its square. Log(Age) denotes log firm age in quarters. Customer Rating denotes a given firm's average Alexa rating (out of five) in 2017Q1 and zero for non-voice-AI firms. Customer Reviews is the mean number of customer reviews (in 10s) for a given firm's Alexa Skills in 2017Q and zero for non-voice-AI firms. The sampling period is 2017Q1 to 2021Q3.

	N	Mean	Median	Stdev.
ROA	30,134	1.62	5.63	24.96
D2AT	30,134	0.57	0.58	0.25
Log(AT)	30,134	7.14	7.23	2.18
Log(AT)^2	30,134	55.80	52.25	31.98
Log(Age)	30,134	2.87	3.04	0.77
Customer Rating	30,134	0.02	0.00	0.24
Customer Reviews (in 10's)	30,134	0.02	0.00	0.27

Table 1 reports summary statistics on our key variables. As shown, the average (median) ROA equals 1.62% (5.63%), debt ratio equals 57% (58%), log assets equals 7.14 (7.23), and log firm age in quarters equals 2.87 (3.04). The average firm has 0.2 comments and a rating of 0.02 out of 5 in the full sample, as we attribute a rating of zero and zero reviews to firms without Alexa Skills.

Average consumer ratings provide important information about customer satisfaction, but their reliability depends on the sample size. The concept of statistical power suggests that, when inferring product qualities, customers should not only look at the average rating of the Skill but also the number of people who rated it, as well as the dispersion of the raters' judgments, which provide important information (Obrecht et al., 2007). For example, an average rating of two stars given by 400 consumers is less noisy than an average rating of two given by five consumers. We, therefore, utilize both of these measures when we analyze the influence of CCPA on businesses.

7 Empirical Strategy

This section presents details on our empirical methodologies. We start with the estimation of homogeneous and heterogeneous treatment effects of the CCPA on sophisticated firms. We then complement these findings with running Double ML techniques that model treatment and estimate heterogeneous treatment effects. To do so, we utilize panel data on U.S. public firms at a quarterly frequency. The main dependent variable of interest is Return on Assets (ROA), which proxies firm profitability.[12] Using the Neyman-Rubin potential outcomes framework, the average treatment effect on the treated (ATT) can be represented as:

$$\begin{aligned} ATT \equiv \delta &= \mathbb{E}[Y_{Post}(1) - Y_{Post}(0) \mid T = 1] \\ &= \mathbb{E}[Y_{Post}(1) \mid T = 1] - \mathbb{E}[Y_{Post}(0) \mid T = 1], \end{aligned} \tag{24}$$

where the second equality follows from the linearity of the expectations operator. Rearranging terms we can write

$$\mathbb{E}[Y_{Post}(1) \mid T = 1] = \delta + \mathbb{E}[Y_{Post}(0) \mid T = 1]. \tag{25}$$

The left-hand side contains a potential outcome variable that can be replaced by its observable counterpart $\mathbb{E}[Y_{Post} \mid T = 1]$ under consistency. As a benchmark counterfactual, we start with assuming $\mathbb{E}[Y_{Post}(0) \mid T = 1] = \alpha_i + \eta_t$, which contains fixed effects for firms α_i and year-quarters η_t. Assuming parallel trends and no pre-trends leads to:

$$\mathbb{E}[Y \mid T] = T \times \{\delta + \alpha_i + \eta_t\} + (1 - T) \times \{\alpha_i + \eta_t\} \tag{26}$$

$$= \delta \times T + \alpha_i + \eta_t. \tag{27}$$

This motivates the estimation of δ with a linear difference-in-differences regression using the two-way fixed effects (TWFE) structure below:

$$Y_{it} = \delta T_{it} + \alpha_i + \eta_t + \epsilon_{it}. \tag{28}$$

[12] We expect private consumer data to improve asset utilization and/or profit margins. ROA allows us to estimate their combined effect and is not influenced by changes in the equity multiplier.

The above model allows us to estimate dynamic treatment effects using panel data on firms i over quarters t and eliminates the risk of control-specific trends.[13] We complement this specification by introducing another counterfactual $\mathbb{E}[Y_{Post}(0) \mid T = 1] = \alpha_i + \delta_t + \gamma \cdot W_{it}$ that contains control variables W_{it} along with heterogeneous treatment effect $\delta(X)$ that measures how CCPA's influence on ROAs of treated units varies across different dimensions. We, therefore, also run regressions on:

$$Y_{it} = \delta(X) \cdot T_{it} + \alpha_i + \eta_t + \gamma \cdot W_{it} + \epsilon_{it}, \tag{29}$$

where $\delta(X) = \delta_0 + \delta_{Reviews} \times$ Reviews $+\delta_{Ratings} \times$ Ratings. The intercept δ_0 measures treatment effects for sophisticated firms with zero reviews and ratings. We think of these firms as having voice-AI products but an insignificant customer base and less in-house data. $\delta_{Reviews}$ and $\delta_{Ratings}$ measure how the influence of the CCCPA on sophisticated firms varies across different proxies of customer base. Importantly, specification (29) allows us to also control for time-varying industry characteristics by introducing industry $\times$ year and industry $\times$ year-quarter fixed effects. These allow us to compare sophisticated firms against firms from their Fama-French-48 industries in a given year or quarter, before and after the introduction of the CCPA.

Treatment effects under unconfoundedness One concern related to the estimation of δ and $\delta(X)$ above is the potential violation of the unconfoundedness assumption. This can make the parallel trends assumption ($\mathbb{E}[Y_{Post}(1) - Y_{Pre}(0) \mid T = 1] = \mathbb{E}[Y_{Post}(1) - Y_{Pre}(0) \mid T = 0]$) unreliable. Although we provide an abundance of evidence on the observed counterparts of the parallel trends assumption, the assumption itself is formally unverifiable because it contains potential outcome variables even after the consistency assumption.

We, therefore, also consider the estimation of average treatment effects (ATE) using a partially linear regression (PLR) model following Robinson (1988) and Chernozhukov et al. (2016, 2017). We model the probability of being a sophisticated firm as a function of confounders indicated as X_i. Confounders affect the treatment variable T_i via the function $f(\cdot)$ and the outcome variable via the

[13] Let $\mathbb{E}[Y_{Post}(1) \mid T = 1]$ and $\mathbb{E}[Y_{Pre}(0) \mid T = 1]$ depend on control variable W. Name the respective coefficients as θ_1 and θ_2. Under this scenario, the true ATT $\delta = \delta^{DD} + (\theta_1 - \theta_2) \cdot W \neq \delta^{DD}$ unless θ's cancel out and our estimate δ^{DD} will end up being a biased estimate of the true δ.

function $\theta(\cdot)$ as follows:

$$T_i = f(X_i) + \eta_i, \tag{30}$$

and

$$\Delta Y_i = \theta(X_i) \cdot T_i + g(W_i) + \epsilon_i. \tag{31}$$

The above specification allows us to estimate a conditional average treatment effect (CATE) of the CCPA in the U.S. To do so, we collapse the time series information first into pre- and post-CCPA periods (i.e., before and after the introduction of the CCPA in 2018) and then take a difference (see, e.g., Bertrand et al. (2004)). The unit of observation i refers to firms, and the outcome variable is once again ROA. W_i contains firm-level control variables that are previously introduced (e.g., pre-CCPA values for debt to assets, log assets, log assets squared, and log age).

$\theta(X_i)$ contains an intercept term, customer ratings as of 2017Q4 along with pre-CCPA values for debt to assets, log assets, log assets squared, sales to working capital, R&D to sales, advertising expenses to sales, and log firm age. We chose these variables to close potential back doors between treatment and our outcome variable, and they allow us to estimate conditional average treatment effects (CATE). The identifiability assumptions of the above specification are $\mathbb{E}[\epsilon_i \mid X_i, W_i] = 0$, $\mathbb{E}[\eta_i \mid X_i] = 0$ for the initial stages of the estimation, and $\mathbb{E}[\epsilon_i \times \eta \mid X_i, W_i] = 0$ for the final stage. We estimate the outcome variable with weighted lasso and sophistication with logistic regression and use five-fold cross-validation.

8 Empirical Findings

This section presents the main empirical findings of our paper. We start with demonstrating the homogeneous and heterogeneous effects of the CCPA on voice-AI firms. These provide empirical support for theoretical predictions laid out in Section 4 and summarized in Figure 2 among other ones.

We start with providing our findings from specification (28) in Table 2. As shown in column 1 of Table 2, we find that a sophisticated U.S. firm attains around 1.59% (t-value = -2.36) lower ROA after the introduction of the CCPA after controlling for Firm $\times$ Year-Quarter fixed effects. Figure

Table 2: **Effects of the CCPA on Sophisticated Firms**

This table reports results from estimating homogeneous and heterogeneous treatment effects of the introduction of the CCPA on sophisticated firms. The dependent variable is Return on Assets. Sophisticated firms are firms with active voice-AI skills on Amazon Alexa, and control firms are firms from the same industries that are not in the Alexa universe. Standard errors are clustered at the Fama-French-48 industry level. Additional details on the empirical strategy are presented in 7. $\star\star\star$, $\star\star$, or $\star$ indicates that the coefficient estimate is significantly different from zero at the 1%, 5%, or 10% level, respectively.

	Return on Assets, %			
	(1)	(2)	(3)	(4)
δ_0	-1.590**	-2.075***	-2.871***	-2.871***
	(-2.36)	(-2.84)	(-3.77)	(-3.74)
$\delta_{Reviews}$		0.630***		-0.004
		(3.63)		(-0.04)
δ_{Rating}			1.104***	1.106***
			(3.59)	(3.76)
Fixed Effects				
Firm	Yes	Yes	Yes	Yes
Industry × Year-Quarter	Yes	Yes	Yes	Yes
Observations	30,134	30,134	30,134	30,134
Adj. R-squared	0.851	0.851	0.851	0.851

8 presents evidence on the dynamics of this effect. As shown in the figure, the reduction in ROA is particularly significant after CCPA became effective in 2020. Columns 2 to 4 of Table 2 present evidence on effect heterogeneity. As shown, sophisticated firms with small customer bases are hit hard by the CCPA. δ_0 varies between -2.075% (t-value = -2.84) and -2.871% (t-value = -3.74) across different specifications. As shown in column 2 (3), a one standard deviation increase in Reviews (Ratings) increases ROA by 0.17% (0.265%).[14] As shown in column 4, we use both of our features in the same regression and show that δ_{Rating} equals 1.106% with a t-stat of 3.76.

Table 3 presents additional findings on effect heterogeneity for treated units. As shown in columns 1-3, we once again find that sophisticated firms with small customer bases are hit hard by the CCPA. A one standard deviation increase in Reviews (Ratings) increases ROA by 0.265% (0.317%).[15] These results are robust to controlling for firm-level characteristics such as debt-to-asset ratio, log book value assets, squared log book value of assets, and log firm age. Columns 4-6 and

[14] $0.27 * 0.63\% = 0.17\%$ and $0.24 * 1.104\% = 0.265\%$.
[15] $0.27 * 0.98\% = 0.265\%$ and $0.24 * 1.32\% = 0.317\%$.

7-9 show that they are also robust to controlling for firm and industry x year and firm and industry x year-quarter fixed effects.

Collectively, Tables 2 and 3 highlight that all sophisticated firms are negatively affected by the CCPA, and sophisticated firms with weak (strong) customer bases experience stronger (weaker) distortionary effects. Based on columns 1 and 2 of Table 3, a treated unit would need to increase its customer ratings by 6.22 ($= 1.984/(1.328 * .24)$) standard deviations or increase its customer reviews by 4.53 ($= 1.195/(0.978 * .27)$) standard deviations in order to counter the distortionary effects of the CCPA. These back-of-the-envelope calculations rely on the identifiability and modelling assumptions but shed some light on the distortionary effects of the CCPA.

One of our theoretical framework predictions is that data regulations can entrench incumbents with in-house data. To provide empirical evidence for this theoretical prediction, we run the TWFE structure from specification (29) to compare sophisticated firms with a big customer base against sophisticated firms with a small customer base in a subsample test. Motivated from our findings from column 1 of Table 2 and columns 3, 6, and 9 of Table 3, we proxy customer base with Rating.

Figure 9 presents dynamic treatment effects estimated from this exercise. As shown in the figure, sophisticated firms with a larger customer base exhibit stronger ROAs after the introduction of the CCPA compared to sophisticated firms with smaller customer bases. The average treatment effect for firms with a big customer base is 4.463% (t-stat= 3.94). The analysis is not tabulated for brevity but is available upon request.

The findings in this section provide empirical evidence consistent with the argument that CCPA hurts voice-AI firms but benefits firms with big customer bases. Such firms exhibit higher ROAs after the adoption of customer privacy laws. One concern related to these findings is the potential violation of the parallel trends assumption, for example, due to firms' anticipation or lobbying efforts to shape the CCPA. Figures 8 and 9 provide evidence on the observed counterparts of the parallel trends assumption and show that we do not have an unnatural experiment. We complement these by running Double ML regressions in which we drop the stronger unconfoundedness assumption as explained in Section 7.

Table 3: **Effect Heterogeneity for Sophisticated Firms**

This table reports results from estimating heterogeneous treatment effects of the introduction of the CCPA on sophisticated firms. The dependent variable is Return on Assets. Sophisticated firms are firms with active voice-AI skills on Amazon Alexa, and control firms are firms from the same industries that are not in the Alexa universe. Standard errors are clustered at the Fama-French-48 industry level. Additional details on variables and empirical strategy are presented in 6 and 7. $\star\star\star$, $\star\star$, or $\star$ indicates that the coefficient estimate is significantly different from zero at the 1%, 5%, or 10% level, respectively.

	Return on Assets,%								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
δ_0	-1.195*	-1.984***	-2.025***	-1.016	-1.713**	-1.776**	-0.965	-1.664**	-1.728**
	(-1.79)	(-3.06)	(-3.18)	(-1.45)	(-2.47)	(-2.61)	(-1.32)	(-2.34)	(-2.47)
$\delta_{Reviews}$	0.978***		0.315	1.067***		0.449	1.064***		0.441
	(3.84)		(0.62)	(3.51)		(0.80)	(3.45)		(0.77)
δ_{Rating}		1.328***	1.154**		1.322***	1.074**		1.328***	1.084**
		(4.16)	(2.57)		(3.88)	(2.14)		(3.97)	(2.18)
γ_{D2AT}	-18.939***	-18.992***	-18.985***	-18.515***	-18.582***	-18.566***	-18.327***	-18.394***	-18.379***
	(-4.17)	(-4.19)	(-4.19)	(-4.00)	(-4.03)	(-4.03)	(-3.97)	(-4.00)	(-4.00)
$\gamma_{Log(AT)}$	21.242***	21.198***	21.212***	20.556***	20.499***	20.522***	20.606***	20.550***	20.573***
	(8.11)	(8.07)	(8.09)	(8.30)	(8.24)	(8.27)	(8.22)	(8.16)	(8.19)
$\gamma_{Log(AT)^2}$	-1.187***	-1.183***	-1.185***	-1.148***	-1.143***	-1.145***	-1.147***	-1.142***	-1.144***
	(-7.17)	(-7.16)	(-7.15)	(-7.25)	(-7.22)	(-7.23)	(-7.15)	(-7.12)	(-7.13)
$\gamma_{Log(Age)}$	1.917	1.899	1.895	0.991	0.980	0.973	0.973	0.961	0.955
	(1.10)	(1.09)	(1.09)	(0.49)	(0.49)	(0.49)	(0.48)	(0.48)	(0.48)
Fixed effects	Firm, Year-Quarter			Firm, Industry x Year			Firm, Industry x Year-Quarter		
Observations	30,134	30,134	30,134	30,134	30,134	30,134	30,134	30,134	30,134
R-squared	0.864	0.864	0.864	0.865	0.865	0.865	0.866	0.866	0.866

Figure 8: **Dynamics of Treatment Effects on Sophisticated Firms**

This figure shows the treatment effects of the CCPA on sophisticated firms in the event time. The first vertical line represents the introduction of the CCPA, and the second vertical line represents its effectiveness date. We present 90% confidence intervals for each point estimate.

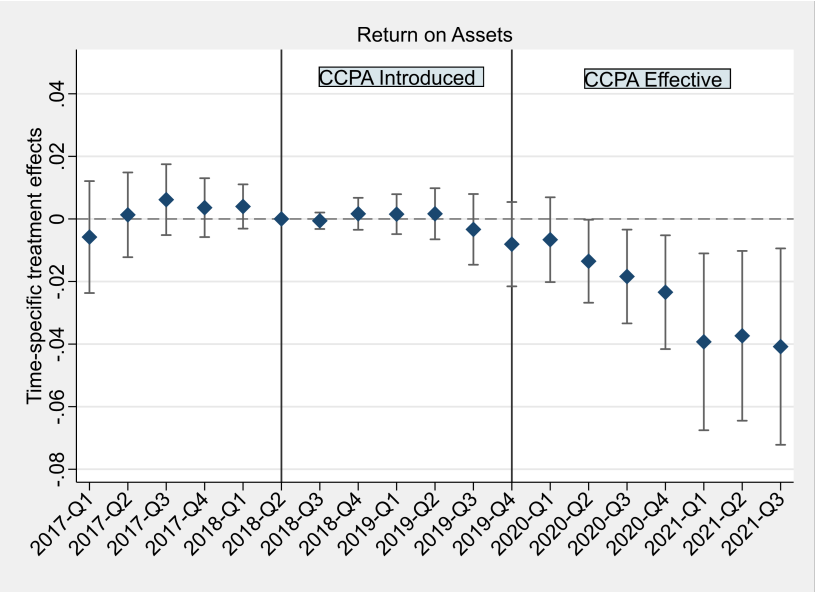


Table 4 presents additional evidence on the economic channel and heterogeneous treatment effects across U.S. firms. As shown, in line with our theoretical predictions and earlier results, technologically-sophisticated firms with small customer bases suffer the most when data trading is affected. On average, firms with lower customer ratings, small firms, and young firms exhibit lower ROAs. The coefficient for customer ratings, θ_{Rating}, varies between 2.522 (t-stat=2.51) and 3.359 (t-stat=2.62) across different specifications.

The above findings collectively reinforce the idea that the adoption of the CCPA affects sophisticated firms with voice-AI products negatively relative to industry peers without voice-AI products. We show this by pinning down negative effects of the CCPA on returns on assets. The negative effect varies across different dimensions. In particular, it is greater for firms with less in-house data, as proxied by low ratings or few reviews. Our findings are robust to a rich array of empirical strategies, control variables, fixed effects structures, and cross-validation.

Table 4: **Effect Heterogeneity Using Double ML**

This table presents results from running a partially linear regression (PLR) model following Robinson (1988) and Chernozhukov et al. (2016, 2017). Confounders X_i affect the treatment variable T_i via function $f(\cdot)$ and the outcome variable via function $\theta(\cdot)$ as:

$$T_i = f(X_i) + \eta_i \tag{32}$$

and

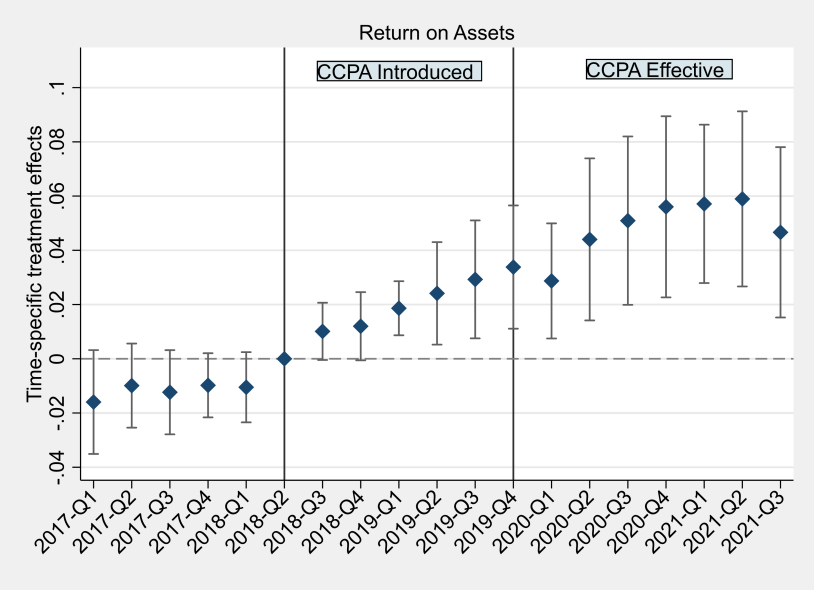
$$\Delta Y_i = \theta(X_i) \cdot T_i + g(W_i) + \epsilon_i. \tag{33}$$

Time series information is collapsed into pre- and post-CCPA periods and we take first differences for the dependent variable, ROA (%). The unit of observation i refers to firms. $\theta(X)$ contains ratings, log assets, log assets squared, sales to working capital, log firm age, debt to assets, R&D to sales, and advertising expenses to sales. W contains firm-level control variables previously used in Table 3 and calculated as pre-CCPA averages. We present $\theta(X)$ coefficients below. Standard errors are bootstrapped. The panel contains 1,586 observations in all columns. $\star\star\star$, $\star\star$, or $\star$ indicates that the coefficient estimate is significantly different from zero at the 1%, 5%, or 10% level, respectively.

	Δ ROA, %		
	(1)	(2)	(3)
$\theta(X)$			
θ_{Rating}	2.522**	3.27***	3.359***
	(2.51)	(2.76)	(2.62)
$\theta_{Log(AT)}$	20.292*	22.677**	25.278**
	(1.89)	(2.07)	(2.26)
$\theta_{Log(AT)^2}$	-1.119*	-1.304**	-1.415**
	(-1.88)	(-2.12)	(-2.28)
$\theta_{SALES2NWC}$	0.015***	0.015***	0.015***
	(10.59)	(10.12)	(7.84)
$\theta_{Log(Age)}$		5.744**	4.655
		(2.44)	(1.57)
θ_{D2AT}			-14.149
			(-1.38)
$\theta_{RD2SALE}$			-0.774*
			(-1.77)
$\theta_{ADV2SALE}$			-25.815
			(-0.75)
Intercept	-90.045	-115.081**	-115.578**
	(-1.94)	(-2.30)	(-2.32)

Figure 9: **Comparison of Sophisticated Firms with Big and Small Customer Bases**

This figure shows the treatment effects of the CCPA on sophisticated firms with a big customer base against sophisticated firms with a small customer base in the event of the CCPA. The first vertical line represents the introduction of the CCPA, and the second vertical line represents its effectiveness date. We present 90% confidence intervals for each point estimate.



9 Conclusion

Quantifying the value of consumer data and consumer privacy is crucial for understanding the rapid period of transformation of the modern data economy. In this project, we exploit the introduction of the CCPA, which limits the amount of consumer data firms can collect and trade but does not affect the amount of data firms can generate in-house to the same extent. We examine a unique and hand-collected data set of conversational-AI firms that rely on consumer data to grow their business. We found that personal consumer data is a competitive advantage for firms in the face of changing privacy regulations and the key to unlocking customer value.

We build a structural model with trade in partially non-rival data. Firms collect data internally or acquire data externally from other firms, allowing them to conduct analysis on the data and generate business-relevant analysis, making labor more productive. We allow for heterogeneity in reliance on data (*sophistication*) and size of a firm's customer base (*ability to generate internal data*). We show that a tightening of privacy regulation may lead to worsening trading frictions, affecting all firms adversely, while sophisticated firms that rely more on data suffer more than unsophisticated

firms. Especially sophisticated firms with small customer bases are hit the hardest. The reason is that such firms find it difficult to substitute missing external data with internally generated data and therefore pay most of the cost of trading frictions.

We confirm these predictions in our empirical analysis. We find that firms with Alexa Skills suffer a larger fall in returns on assets than firms without Alexa Skills. We find that within the group of firms with Alexa Skills, those with higher ratings and a larger number of customer reviews suffer less than firms with low reviews and few customer reviews. These results are robust to firm, year-quarter, and industry x year-quarter fixed effects. We provide additional evidence using double-machine learning techniques to highlight the heterogeneity and economic mechanisms through which these negative effects of CCPA manifest. In particular, we find that firms with low customer ratings that are young and small suffer the most.

References

Acemoglu, Daron, "Harms of AI," Technical Report, National Bureau of Economic Research 2021.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman, "The Economics of Privacy," *Journal of Economic Literature*, 6 2016, *54* (2), 442–492.

Aridor, Guy, Yeon-Koo Che, and Tobias Salz, "The Effect of Privacy Regulation on the Data Industry: Empirical Evidence from GDPR," NBER Working Papers 26900, National Bureau of Economic Research, Inc 2020.

Babina, Tania, Anastassia Fedyk, Alex Xi He, and James W. Hodson, "Artificial Intelligence, Firm Growth, and Industry Concentration," Technical Report 2021.

_ , Greg Buchak, and Will Gornall, "Customer Data Access and Fintech Entry: Early Evidence from Open Banking," SSRN Working Paper 4071214 March 2022.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan, "The economics of social data," Technical Report 2020.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-In-Differences Estimates?," *The Quarterly Journal of Economics*, 2004, *119* (1), 249–275.

Bessen, James, Stephen Impink, Lydia Reichensperger, and Robert Seamans, "GDPR and the Importance of Data to AI Startups," SSRN Working Paper 3576714 2020.

Campbell, James, Avi Goldfarb, and Catherine Tucker, "Privacy Regulation and Market Structure," *Journal of Economics & Management Strategy*, March 2015, *24* (1), 47–73.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey, "Double Machine Learning for Treatment and Causal Parameters," *ArXiv e-prints*, 2016.

_ , _ , _ , _ , _ , and _ , "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review: Papers Proceedings*, 2017, *107* (5), 261–265.

Cong, Lin William, Danxia Xie, and Longtian Zhang, "Knowledge Accumulation, Privacy, and Growth in a Data Economy," *Management Science*, 2021.

de Matos, Miguel Godinho and Idris Adjerid, "Consumer Consent and Firm Targeting After GDPR: The Case of a Large Telecom Provider," *Management Science*, 2021.

Eeckhout, Jan and Laura Veldkamp, "Data and Market Power," Columbia University WP Sep 2021.

Farboodi, Maryam, Roxana Mihet, Thomas Philippon, and Laura Veldkamp, "Big Data and Firm Dynamics," *AEA Papers and Proceedings*, May 2019, *109*, 38–42.

Hagiu, Andrei and Julian Wright, "Data-enabled learning, network effects and competitive advantage," NUS WP 2020.

Hoberg, Gerard and Gordon M. Phillips, "Scope, Scale and Competition: The 21st Century Firm," SSRN Working Paper 3746660 2021.

Jia, Jian, Ginger Zhe Jin, and Liad Wagman, "The Short-Run Effects of GDPR on Technology Venture Investment," *Marketing Science*, Mar 2021, *40* (4), 661–684.

Jones, Charles I and Christopher Tonetti, "Nonrivalry and the Economics of Data," *American Economic Review*, 2020, *110* (9), 2819–58.

Khansa, Lara, Deborah Cook, Tabitha James, and Olga Bruyaka, "Impact of HIPAA provisions on the stock market value of healthcare institutions, and information security and other information technology firms," *Computers & Security*, Sep 2012, *31*, 750–770.

Obrecht, Natalie A, Gretchen B Chapman, and Rochel Gelman, "Intuitivet tests: Lay use of statistical information," *Psychonomic bulletin & review*, 2007, *14* (6), 1147–1152.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer, "European Privacy Law and Global Markets for Data," Technical Report 2021.

Robinson, P. M., "Root-N-Consistent Semiparametric Regression," *Econometrica*, 1988, *56* (4), 931–954.

Rydning, David Reinsel-John Gantz-John, "The digitization of the world from edge to core," *Framingham: International Data Corporation*, 2018, p. 16.

Vesset, Dan and Jebin George, "Worldwide Big Data and Analytics Spending Guide," Technical Report, IDC 2021.

Vives, Xavier and Zhiqiang Ye, "Information Technology and Bank Competition," Technical Report 2021.