Does the Market Value Attention to Data Privacy? Evidence from U.S.-listed Firms under the GDPR

Farzam Boroomand, University of Minnesota, farzam@umn.edu

Aija Leiponen, Cornell University, aija.leiponen@cornell.edu

Gurneeta Vasudeva, University of Minnesota, gurneeta@umn.edu

Abstract

Despite the growing interest in understanding the big data revolution and its implications for business, our knowledge has remained limited regarding the evolution of firms' data practices and their strategic and economic implications. In this study, we focus on firms' practices related to data privacy. We quantify firms' attention to data privacy via a text analysis of U.S. publicly listed firms' annual reports. We show that firms' attention to data privacy responds to changes in regulatory, normative, and competitive pressures and has substantial financial implications. U.S. firms increased attention to data privacy as a result of their exposure to the European GDPR regulation and data breaches in their industry. Firms responded similarly if their competitive peers were exposed to the regulation or a breach. The increased attention to data privacy contributed to a reduced market valuation. These findings suggest that firms that respond to regulatory, normative, and competitive threats related to data privacy expend managerial attention and resources on data practices and technologies, but such practices do not allow them to differentiate themselves in the market. However, firms with technological capabilities in data analytics (machine learning patents) and those with a market presence in weak data privacy regimes were able to mitigate the adverse effects of stringent demands for data privacy. We discuss the broader managerial and policy implications of these findings as well as the potential avenues for future research.

Key words: big data, data privacy, GDPR, data breaches, machine learning, innovation

1. Introduction

In November 2020, the Chinese financial-technology company, Ant Group, was valued in advance of its Initial Public Offering on the Shanghai Stock Exchange at USD 34.5 billion (Statista 2021). Analysts attributed this staggering market valuation in large measure to Ant's "most valuable assets—its data**,** derived from billions of consumer transactions" (Forbes 2021). While data analytics have become the basis of competitive advantage for some of the most valuable platform-based firms such as Ant, Alibaba, Apple, Uber, Facebook and Google, even less technologically sophisticated firms in traditional industries such as banking and finance, retail, and healthcare are adopting data-centric business models (Brynjolfsson and McElheran 2016, Tambe 2014, Tambe and Hitt 2012). Meanwhile, advances in enabling technologies for data analytics such as artificial intelligence, machine learning, and the internet of things are spurring new research agendas aimed at harnessing the data revolution (National Science Foundation 2019). However, comparatively less is known about firms' investments and practices related to data resources, in part because it is challenging to quantify or value data assets. In this paper we develop a new text-based measure of firms' attention to data practises. We examine how regulation, competition, and industry norms influence attention to data practices and explore their implications for market performance.

As much of the tremendous market value created by firms' data assets depends on personal data, a particularly vexing issue related to data practices is the tension between the lack of clear property rights to data and the inherent "inalienability" of information in the data vis-a-vis the data subject (Thomas et al. 2022). The value of any instance of data depends on its tight connection to an individual, organization, or phenomenon. For example, the value of health data or shopping data arises from its ability to predict individual users' behavior. However, such data cannot be owned; it can only be partially controlled by by secrecy combined with a patchwork of regulations and contractual agreements related to its sharing and utilization. The data subject may consider the data private or confidential and prefer to prevent its dissemination, even though they may not have any rights to prevent the sharing. Markets for certain types of data can thus be ethically repugnant yet legal (Koutroumpis, et al. 2020). In many contexts, the rights

and regulations related to digital privacy are contentious and in flux, and data sharing is subject to significant ethical, regulatory, and competitive uncertainty.

Firms' data practices are also challenged by increasingly sophisticated hackers. Data breaches are common and affect all industries where valuable data resources are processed. According to the Identity Theft Resource Center, 1108 data breaches were reported in the United States in 2020, affecting billions of personal records.[1] Even an average data breach event cost the affected firm millions of U.S. dollars and was likely to destroy substantial market value of publicly listed firms.[2]

The foregoing challenges of the big data revolution propelled an institutional intervention by the European Union resulting in the General Data Protection Regulation (GDPR) which aimed at setting strict standards to govern firms' data privacy and security practices. GDPR protects data belonging to E.U. citizens and residents, but the law also applies to organizations that handle such data whether they were E.U.-based organizations or not.[3] At the time of its implementation, GDPR was probably the most stringent data regulation in the world, and its economic effects were documented by several studies (see e.g. Jia et al., 2021; and Peukert et al. forthcoming).

In this study, we illuminate U.S. firms' attempts to navigate the evolving market for data and associated risks related to data privacy. We define data practices as how firms manage the vast amounts of data they analyze and process (Thomas et al. 2021). The growing demand for data privacy may create opportunities for firms to devise data practices to differentiate themselves from competitors as providers of desirable privacy goods and services. However, innovative data practices come with implementation costs and potential revenue implications. For example, when Apple invested in new software on its mobile devices that placed data privacy controls in the hands of users, the resulting privacy restrictions

[1] https://www.securitymagazine.com/articles/96667-the-top-data-breaches-of-2021 retrieved on 6 January, 2022.
[2] https://www.ibm.com/security/data-breach retrieved on 6 January, 2022.
[3] https://gdpr.eu/companies-outside-of-europe/ retrieved on 6 January, 2022.

limited the firm's access to users' data that was critical to tailored digital advertising on its platform.[4] Thus, the provision of privacy goods may hurt the revenue-generation potential of digital platforms.

Despite the increasingly central strategic role of data resources, the availability of research data for analyzing firms' data practices is limited. To overcome this informational challenge, we focus on firms' *attention to* data practices. We assume that, although attention to data risks does not guarantee action, data strategies cannot come about in the absence of attention (Ocasio, 1997). Our study employs text-based analysis and modeling to discover and quantify attention to data privacy and security by publicly-traded U.S. firms. We apply topic modeling based on Latent Dirichlet Allocation (LDA) (Blei et al. 2003, Hannigan et al. 2019) to analyze the 10-K forms filed by publicly traded companies in various industry and technology sectors between 2000 and 2018. The topics that emerge from the analysis vary by firm and by industry and meaningfully reflect differences in firms' approaches to exploiting data in their digital transformation.

We argue that U.S. firms that are directly subject to GDPR owing to their presence in the E.U. market and those U.S. firms indirectly facing competitive pressure from peers operating in the E.U. are likely to devote more managerial attention and resources towards devising data practices to address privacy concerns. We also examine the normative role of industry-level data breaches in driving attention to data privacy. Once an industry peer is subject to a data breach, close competitors may be motivated to pay more attention to practices related to data privacy risks. We explore the heterogeneity across firms in terms of their disclosed attention to data privacy risks, and the market value of this disclosure.

Clearly disclosing their attention to data risks becomes even more important when firms' investments in information technology involve intangible assets. By making these assets more visible and measurable, firms may be able to increase their market valuation (Brynjolfsson et al. 2002, Hall 2001, Saunders and Tambe 2015).

[4] https://www.wsj.com/articles/p-g-worked-with-china-trade-group-on-tech-to-sidestep-apple-privacy-rules-11617902840?mod=djemalertNEWS retrieved on 6 January, 2022.

We find that GDPR implementation created not only regulatory pressures for U.S. firms conducting business in the E.U., but also incentives for firms to opt in and increase attention to data privacy because of competitive pressure from peers active in the E.U.. Firms also responded to normative pressures created by data breaches of their peers by increasing their attention to data privacy issues. Furthermore, we find that the increased attention to data privacy had negative implications for firms' market value. It appears that the costs associated with enhanced attention to data privacy exceed any benefits from differentiation in the marketplace via higher prices or larger market shares (Koutroumpis, Leiponen and L. D. Thomas 2020, Leiblein et al. 2018, Villalonga 2004). Increased attention to data privacy, thus, had adverse effects on firms' financial performance. However, we find that firms that had made substantial investments in machine-learning technologies enabling data analytics, and those with a market presence in weak data privacy regimes were better able to shield themselves from the detrimental effects of having to pay attention to data privacy, and thereby mitigate the adverse market performance effects.

Our study makes three main contributions to the literature on the strategic and economic implications of digitization for firms in a broad range of industries. First, using topic modeling we discover and quantify firms' attention to various types of data risks in a large sample of publicly-listed firms. The emerging topics highlight that provision of data privacy is a distinct data practice. Second, our findings suggest that U.S. firms respond to regulatory, competitive, and normative pressures by increasing attention to risks related to data privacy. Third, attention to data privacy and associated response strategies may not allow firms to reap higher market valuations. Our findings suggest that the institutional forces propelling firms towards internalizing the cost of data privacy may not be rewarded by consumers and investors. Nevertheless, some firms may be able to mitigate the adverse impact through investments in new technologies and market access.

2. Hypotheses

Although digital forms of data and information systems have existed for decades, research on the strategic implications of data assets is extremely limited. Following Thomas et al. (2020), we define data as *codified observations fixed in a tangible medium*. These observations have not yet been significantly processed or manipulated. Data can be obtained from social interactions, laboratory experiments, production processes, or other types of measurements of the physical or digital environment compiled by humans or machines (Lycett 2013, Uhlir and Cohen 2011). Unstructured data can be transformed into resources that have grouping, relatedness, and purpose (Borgman 2012). Such structured databases can be analyzed using software programs to generate information and insights (Chebli et al. 2015).

Data resources and associated practices may enable the firm to improve products and services (e.g. using customer insights to inform innovation) or production and logistical processes (e.g. using operational analytics to identify productivity bottlenecks), or the data can be monetized outside the firm to generate such insights for external partners (Wu et al. 2020). Data practices may thus have strategic implications for the long-term performance of the firm by facilitating durable differentiation or cost leadership in the market vis-a-vis complementors and competitors (cf. Leiblein et al. 2018).

2.1 Regulatory drivers of firms' attention to data privacy

Although data resources can be a persistent source of value for firms, such information advantages also present challenges that may violate the privacy and security of entities whose data are collected (Barrett 2018, Khan 2019). Technology platforms monetize behavioral data directly through various trading or licensing arrangements and indirectly through targeted advertising. The platform design and proprietary algorithms make these data practices difficult to detect. Nevertheless, the tangled web of data sharing in interdependent digital ecosystems can result in direct and indirect harms to users, while regulatory weaknesses and lax enforcement perpetuate such data privacy risks.

The GDPR regulation in the E.U. recognizes these concerns and imposes tough data privacy and security requirements on firms that target or collect personal data of individuals residing in the EU. These tough standards contrast with the U.S. privacy laws in normative and structural terms. Normatively, as Barrett (2018, p. 1065) observes, "U.S. privacy protections are hobbled by U.S. privacy law's predominant objective of facilitating a robust environment for technological innovation and philosophically weakened by a conception of privacy as a good to be traded away, rather than a right to be protected." From a structural standpoint, although an array of sector-specific state and federal statutes have established limitations for data collection and use, the narrow scope of these regulations leaves open vast avenues for digital harms caused by data manipulation or discrimination. Critics argue that the practice of "notice and choice" only creates an illusion of compliance while imposing a high cognitive burden that requires users or complementors to evaluate whether firms' data practices are unfair or deceptive. Consequently, in the U.S., firms can largely exploit the data they have collected so long as they are "transparent" about their policies. Such institutional differences between the U.S. and the E.U. influence the data strategies and associated costs of regulatory compliance for U.S. firms.

In particular, the GDPR strengthens individuals' privacy rights by creating strict personal data management requirements and substantial enforcement mechanisms. The law compels U.S. firms conducting business in the E.U. to increase attention to data privacy and security and raises the bar on how U.S. firms collect data and generate value from it. Even when firms are not exposed to the regulatory mandate, they may voluntarily adopt it when their competitive peer firms experience regulatory mandates that influence consumer expectations in their market (Aragòn-Correa et al. 2019, Vasudeva et al. 2018). Yet, as Peukert et al. (forthcoming, p. 1) observe, "while the GDPR may apply to firms in the EU catering to consumers outside the EU and to firms outside the EU catering to consumers in the EU, it does not, *de jure*, apply to firms outside the EU catering to consumers outside the EU."

This type of proactive adoption may also arise from the expectation of regulatory spillovers across jurisdictional boundaries or the firms' expectation to enter the market with the regulatory mandate in the foreseeable future (Bessen et al. 2020, Fremeth and Shaver 2014, York et al. 2018). For example,

following the implementation of GDPR, U.S. firms *without* market presence in the EU may choose to voluntarily adopt parts of regulations set by GDPR as they may foresee the passage of state-level data privacy legislations such as California Consumer Privacy Act or federal level regulations such as the Data Care Act 2018 bill seeking to establish duties of care and confidentiality for digital platforms much in the same way as the GDPR. Firms subject to the GDPR mandate may seek to catalyze similar regulations in the U.S. Consequently, firms may choose to opt in for competitive reasons even when they do not themselves operate in a jurisdiction where GDPR is mandatory.

In addition to the regulatory pressure to enhance privacy, firms encounter normative pressures related to privacy. Specifically, data breaches may change consumers' expectations regarding data privacy and security. With the frequent occurrence of data breaches in the U.S., these expectations may evolve significantly over time (McAfee Labs 2016, Verizon 2017). A data breach involves "the compromise of confidentiality, integrity, or availability of data or information technology (IT) assets that are responsible for the creation, storage, processing, transport and safeguarding of data assets" (Benaroch and Chernobai 2017, p. 1). In certain instances, the breached data are leaked, usually by a criminal entity seeking to sell the data in the black market. When a breach takes place, the targeted firm tends to tarnish its reputation and lose a substantial amount of business, and thus market value (Bundy et al. 2017), but the competitive peers of the targeted firm may also suffer. As consumers lower expectations about privacy practices of the breach target, rivals' privacy practices may also be scrutinized. As a result, rivals may pre-emptively enhance their attention to privacy practices and increase the associated communication to investors in response to a consumer data breach (Say and Vasudeva 2020).

Our conceptualization of firms' data privacy practices draws on the literature on managerial cognition that emphasizes managerial attention as a prerequisite to managerial action (Ocasio 1997). Whereas attention does not guarantee action, it is unlikely that strategically meaningful action takes place without any managerial attention. This perspective allows us to view firms' annual reports as reflecting managerial attention that is potentially conducive to strategic action (Bao and Datta 2014, Ocasio 1997,

Ocasio and Joseph 2018). Therefore, we expect that institutional, competitive, and normative pressures for compliance will direct managerial attention toward the risks and opportunities related to data privacy.

> ***H1a:*** *Following the issuance of the General Data Protection Regulation (GDPR), U.S. firms with a greater E.U. market presence will pay more attention to data privacy.*
>
> ***H1b****: Competitive peers' market presence in the E.U. will increase U.S firms' attention to data privacy post GDPR.*
>
> ***H1c****: Data breaches experienced by competitive peers will increase U.S firms' attention to data privacy.*

2.2 Market valuation outcomes of firms' attention to data privacy

In the evolving institutional and competitive context surrounding data privacy practices, many U.S. firms consider adopting strategies that safeguard against data privacy hazards. Such strategic renewal aimed at long-term value creation and survival requires doing away with many existing practices of "permissionless innovation" (Barrett 2018) that contradict the emerging norm of data privacy as an inherent right. When privacy is viewed as a right rather than a risk that consumers must take, it is likely to be associated with enhanced expectations about the service quality that must be developed and incorporated into product and service features.

However, the provision of privacy features can lead to substantial operational and technical costs. Such costs arise because firms not only need to fulfill their own fiduciary responsibilities by instituting new practices to protect and control access to data, but also assume diffused responsibility for a wide variety of complementors and third-party services in their ecosystem (Puranam et al. 2012, Kretschmer et al. 2020). To establish control or authority over the data practices of ecosystem participants may require investments in redesigning the organizational architecture, governance procedures, and incentive structures to institute stricter data security protocols and prevent unauthorized use of data. For example, Peukert et al. (forthcoming) find that, after GDPR, websites of E.U. based companies significantly reduced third-party tracking "cookies" that interfere with users' privacy. Firms may thus need to invest in software, skills, practices, and data management techniques to address data privacy concerns.

On one hand, the large intangible component of these new resources should increase firms' market value (Kogut and Zander 1992, Peteraf 1993, Barney 1996). For instance, Villalonga (2004) found that intangible assets captured by Tobin's q were strongly correlated with firms' sustained profitability. On the other hand, firms' ability to increase prices and reap the benefits of differentiation based on their investments in intangibles may be limited, at least in the short run: abiding by the regulation would no longer permit firms to employ personal data to generate additional revenues without explicit opt-in by users. However, collecting and managing opt-in can involve expenditures greater than consumers' willingness-to-pay for improved privacy (Godinho de Matos and Adjerid 2021, Goldfarb and Tucker 2011, Miller and Tucker 2009).

Consequently, in many consumer-oriented industries, a stricter data privacy regime can severely curtail data-driven targeted advertising and brand building that is crucial for revenues, while adding technical costs to institute new practices to more securely collect, access, and exploit consumers' data. The net effect of higher costs and reduced revenues resulting from stricter data privacy practices may weaken financial performance outcomes, although such practices may shield firms from the potentially catastrophic events associated with data breaches, legal battles, regulatory fines, and other organizational and reputational damages associated with data privacy violations.

Notwithstanding these expanded costs and less optimistic growth prospects, customers and investors may reward improved data privacy performance, thereby allowing a firm to better differentiate itself from rivals. Yet, while privacy violations have a negative impact on consumer trust and negatively impact firms' financial performance (Martin 2020), adhering to stricter standards of data privacy and security may become a norm that goes unrewarded. In other words, while regulatory violations or breaches of data privacy will most certainly cause hefty financial and reputational damage for firms, absent such events, enhanced data privacy may not confer a differential pricing power to increase revenues. In short, although the impact of enhanced attention to data privacy on pricing or differentiation-based value creation is ambiguous, the cost effects are unequivocal, resulting in a potential decline in profitability in the short run.

***H2a**: Increased attention to data privacy will decrease firms' market valuation.*

2.3 The Contingent Role of Technological and Geographical Factors

Despite the foregoing prediction of the negative performance effects of enhanced attention to data privacy, we suggest that two countervailing mechanisms pertaining to key strategic factors could mitigate the decline in performance. First, firms increasing their attention to data privacy would also need to enhance attention towards building software and hardware capabilities for exploiting their data resources more efficiently (Brynjolfsson and McElheran 2016). In particular, firms with accumulated patented technologies for managing and analysing data may hold advantages in terms of cost-effectively implementing data privacy protections that differentiate them from competitors and win the consumers' trust. For example, machine learning technologies may allow firms to draw inferences from limited datasets and thereby compensate for the constrained access to data under privacy restrictions. These technologies may also allow firms to implement enhanced data privacy practices more efficiently. For this reason, it is plausible for large technology firms such as Facebook and Google to even extend their market leadership under GDPR[5]. For example, IBM has developed a patent for "[p]rivacy violation detection of a mobile application program" that feeds learnable features of a mobile app "into a machine-learning-based classification algorithm" to detect whether a "mobile application program includes one or more permissions for accessing unauthorized privacy data of a mobile application user" (Ferrara et al. 2020). Such an invention enables the firm to innovate with personal data, while staying compliant with privacy requirements. Therefore, technological capabilities embodied in such patents may serve as evidence to the market that a firm can efficiently address the demands of the privacy regime.

Second, when firms depend on external data resources such as users' data for their machine-learning technology and product development, firms with a market presence in weak data privacy regimes may continue to build such capabilities while incurring lower costs of innovation. Weak data privacy regimes allow collection and processing of personal data with minimal restrictions thereby compensating

[5] https://www.wsj.com/articles/gdpr-has-been-a-boon-for-google-and-facebook-11560789219 retrieved on 6 January, 2022.

for the data constraints imposed by GDPR. Based on the personal information security standards in China, for instance, user device data, such as the device start-up time, model, time zone, country, language and IP address are not counted as 'personal information[6].' To exploit this type of personal information, major firms such as the U.S. consumer goods giant Procter & Gamble are reportedly forging alliances with state-backed Chinese trade groups and technology firms to devise technological solutions that circumvent the data privacy practices that the GDPR seeks to establish[7]. One such technological solution tests an algorithm to track iPhone users for targeted advertising in a way that Apple is seeking to prevent.

Taken together, firms' technological capabilities and geographical presence in specific markets could represent valuable strategic factors that offset the costs associated with data privacy and improve firms' market value.

> ***H2b****: Technological capabilities in digitization mitigate the negative effect of attention to data privacy on market valuation in H2a.*
> ***H2c****: Market presence in weak data privacy regimes mitigates the negative effect of attention to data privacy on market valuation in H2a.*

3. Measuring attention to data practices from 10-K reports

3.1 Overview

We develop an approach to discover and quantify firms' data privacy strategies from their 10-K reports. 10-K reports are annual disclosures required by the U.S. Securities and Exchange Commission (SEC) where publicly traded firms describe their business and the risks associated with it (Bao and Datta 2014, Hoberg and Phillips 2010). We base our methodology on Item 1 (i.e., business overview) and Item 1A (i.e., risk factors) of these reports. These sections provide detailed information about firms' "business description" and the "most significant factors that make [firms'] offering speculative or risky" (Bao and Datta 2014, Hoberg and Phillips 2010, Security Exchange Commission 2005). We discover and quantify

[6] https://www.wsj.com/articles/p-g-worked-with-china-trade-group-on-tech-to-sidestep-apple-privacy-rules-11617902840 retrieved on 6 January, 2022.
[7] https://www.wsj.com/articles/p-g-worked-with-china-trade-group-on-tech-to-sidestep-apple-privacy-rules-11617902840

firms' attention to data practices from the textual data provided in 10-K annual reports using topic modeling techniques and particularly the Latent Dirichlet Allocation (LDA) model. We also compare our measurements of data strategies using LDA against measurements of those data strategies developed using word counts. Content analysis techniques such as word counts have previously been used to capture firms' attention to certain aspects of their business (e.g., Eggers and Kaplan 2009, Gamache et al. 2014). Figure 1 presents an overview of the methodological approach to discover and quantify firms' attention to data practices.

*** Insert Figure 1 about here ***

3.2 Textual Data and Sample Description

To prepare our sample, we first collected ticker symbols of publicly traded U.S. firms across four different industries identified by two-digit NAICS codes from COMPUSTAT: 1) NAICS sector 51 – information, 2) NAICS sector 52 – finance and insurance, 3) NAICS sectors 44 and 45 – retail, and NAICS sector 62 – healthcare. Our choice of industries was based on the extent to which industries were likely to employ "big data" for their business operations and strategic decision-making. We referred to the available literature to make our choices (e.g., Abbasi et al. 2016, Kitchin 2014). As part of this step, we sampled 1,259 publicly-traded U.S. firms across the previously mentioned industries.

We then downloaded 10-K annual reports of the sampled firms filed with SEC between 2006 and 2019. The starting year of our sample follows the decision by SEC in 2005 which required all publicly traded firms to include a separate section (Item 1A – risk factors) in their 10-K annual report to identify the "the most significant factors that make [their] offering speculative or risky" (Security Exchange Commission 2005). We downloaded 10-K annual reports in HTML format. A total of 14,707 reports were originally downloaded. We used available tags in the HMTL code to automatically parse out the downloaded reports. Particularly, we used tags in the reports' 'table of content' to extract Item 1 and Item 1A from the reports. Not all reports had an HTML structure that could be automatically parsed out. Particularly, most reports filed prior to 2010 did not have an HTML structure. Accordingly, we limited

our sample to 11,579 reports filed between 2010 and 2019. We were able to automatically parse out Item 1 and Item 1A from 8,707 10-K annual reports.

We then split extracted items into separate sentences. We retained all sentences that contained the word "data" and discarded the rest. We preserved the metadata about the report from which each sentence was extracted (i.e., company's ticker symbol, report's filing year, company's NAICS code, and the URL to the report). We constructed a corpus of documents to train an LDA model such that each document in the corpus was a sentence containing the word "data" extracted either from Item 1 or Item 1A of the 10-K annual reports in our sample. We extracted 68,962 sentences from Item 1 and 60,441 sentences from Item 1A. Sentences extracted from Item 1 and Item 1A had average lengths of 35.22 words and 42.30, respectively. The length of sentences is particularly important since research on LDA models has demonstrated that the technique underperforms if the documents are too short (Tang et al. 2014). Prior management research has also successfully used documents of similar length to ours such as employee reviews on Glassdoor (Corritore et al. 2019) and tweets (Hannigan et al. 2019). In summary, we extracted 129,133 sentences with the average length of 38.53 words from Item 1 and Item 1A of 8,707 10-K annual reports.

We then cleaned the extracted sentences by 1) removing the stop words, i.e., words that "serve a less important role in meaning construction" (Hannigan et al. 2019, p. 592), 2) removing the punctuation characters and digits, 3) stemming – i.e., the "conversion of text segments (words) to their root word forms" and lemmatizing – i.e., "transforming a word into its dictionary form" (Hannigan et al. 2019, p. 592). We also used Gensim Phrases command to automatically detect bigrams that appeared more than 20 times across documents. Bigrams are two-word units rather than individual words that often appear together. Examples of bigrams that were detected in our documents include *third party, security breach,* and *data protection.*

3.3 Discovering data privacy strategies with the LDA topic model

After constructing the corpus of documents that contained relevant textual data, we used topic modeling to discover and quantify data privacy strategies. Topic modeling has been used as a type of statistical modeling technique to discover a set of topics that describe a collection of documents. Particularly, we used latent Dirichlet allocation (LDA) topic model to simultaneously discover and quantify the topics in a set of 10-K reports. LDA was originally developed by Blei et al (2003) as a natural language processing (NLP) technique with the goal of "find[ing] short descriptions of the members of a collection that enable efficient processing of large collections" (Hannigan et al. 2019). LDA has emerged as the most popular topic modeling technique across various disciplines including management scholarship (Bao and Datta 2014, Hannigan et al. 2019).

The LDA model is a bag-of-words method which assumes exchangeability of words within a document (i.e., order of words does not matter). The exchangeability assumption implies that single words (or higher order word combinations such as bigrams) in a document have a mixture distribution. This assertion has been proven mathematically in prior research (De Finetti 1990). The probability distribution of a mixture distribution is a function of another random variable which has its own probability distribution. In the case of LDA model, it is assumed that words have a Poisson distribution across topics and topics have a Dirichlet distribution across documents.

The LDA model requires the researcher to select the number of topics *a priori*. Since we are interested in discovering data strategies, we require the discovered topics to be "clear and well-bounded" (Hannigan et al. 2019). Thus, we followed the methodology proposed by Hannigan et al. (2019) to maximize the average coherence score across topics. Accordingly, we generated 19 different LDA models with number of topics ranging from 2 to 20 with step of 1 using Python Gensim library. We then computed the coherence score for each of the LDA models and selected the number of topics such that we maximized topic coherence. Our final topic model has seven topics and demonstrated maximum coherence score of 0.555. The top 10 terms in each of the topics are displayed in Table 1.

*** Insert Table 1 about here ***

After training the LDA model we examined the discovered topics for their meaning. In addition to the LDA model with seven topics we also examined other LDA models that we had trained in the previous step to find out if any other LDA model uncovers topics that are more interpretable (despite the fact that other LDA models had lower coherence score). Our examination of different LDA models with different topics confirmed that some similar topics existed across all trained LDA models. For example, topics related to data security, data privacy, and third-party data service providers existed across all trained models. Since topics uncovered by our selected LDA model were well interpretable and the coherence score of the selected model was the highest, we proceeded with these topics.

Before using the topics learned by the topic model, the topics need to be labeled and validated. Although there are algorithms that can automatically label the generated topics, these algorithms often perform poorly when domain-specific knowledge is required (Bao and Datta 2014). Thus, we designed a manual labeling procedure to make use of human experts' domain-specific knowledge based on the available literature on big data (e.g., Abbasi et al. 2016, Chen et al. 2012). For each of the seven topics, we used the top 10 words that represent each topic and compared them against documents that were highly associated with each topic. If a topic was dominant in a document, the meaning of that document and the top 10 words in the topic, allowed us to capture the meaning behind the document, to understand the logic behind the topic assignment, and to come up with meaningful labels for each topic. Accordingly, we labeled the topics as presented in Figure 1.

3.4 Constructing measures from the topic model

We used the distribution of topics in any given document (i.e., θ_s) to develop a firm-level measurement of the discovered topics. In our selected topic model with seven topics, θ_s is a vector $[\theta_{s,1}, \theta_{s,2}, \dots, \theta_{s,7}]$ that represents the topic proportions for document s. Each document in our sample is a *sentence* extracted from a 10-K annual report of a firm that contains the word "*data*". Thus, $\theta_{s,k}$is the proportion of topic k in sentence s. To construct a firm-level measure of attention to data strategies for a given annual report, we aggregated θ_s vectors across *s* sentences that were extracted from the focal report. Assuming $N_{i,t}$

sentences were extracted from a 10-K annual report of firm *i* in year *t* we have a set of $\{\theta_{s_1}, \dots, \theta_{s_k}, \dots, \theta_{s_{N_{i,t}}}\}$ vectors each representing the topic proportions for a sentence s in the report. To construct a firm-level measure of data strategies for firm *i* in year *t* we aggregated θ_s vectors such that:

$$DS_{i,t} = \sum_{k=1}^{N_{i,t}} \theta_{s_k} = \left[\sum_{k=1}^{N_{i,t}} \theta_{s,1}, \dots, \sum_{k=1}^{N_{i,t}} \theta_{s,7}\right] = \left[\Theta_{i,t,1}, \dots, \Theta_{i,t,7}\right]$$

Thus, we have a vector of $DS_{i,t} = \left[\Theta_{i,t,1}, \dots, \Theta_{i,t,7}\right]$ for each firm-year that we use as a measure of firms' attention to data strategies that year.

3.5 Constructing measures from a dictionary approach

As a validation step for our LDA based measurements, we also constructed complementary dictionary-based measurements for attention to data privacy and attention to data security variables. These two variables are particularly important for the empirical context of GDPR, and thus we focus on developing a dictionary-based measure only for these variables. For our dictionary-based measurements, we use a simple word count because of their greater transparency. We do this in line with previous research as word counts have been previously used to capture firms' attention to certain aspects of their business (e.g., Eggers and Kaplan 2009, Gamache et al. 2014). To construct our dictionaries for word counts, we reviewed all words in the dictionary of the corpus of documents after filtering out stop words, and removing words that appeared in more than 90%[8] of all documents, and those that appeared in less than 5% of all documents as a common practice in training LDA models (Hannigan et al. 2019). We then extracted security and privacy related words. Our dictionary for security words and bigrams includes {*"security", "breach", "cyber", "information_security", "security_breach", "unauthorized_access", "cyber_attack", "security_measure", "security_standard"*}, and our dictionary for privacy words and bigrams includes {"*privacy", "personal_data", "personal_information", "data_privacy", "data_regulation", "privacy_regulation"*}. Note that terms with an underline represent bigrams. We then

[8] We maintained the word *data* despite its appearance in all documents in our sample because of the focus of our study on data strategies.

constructed dictionary-based measurements of attention to data privacy and attention to data security by simply counting the words representing each construct in the sentences extracted from each 10-K annual report. In the following section, we evaluate our measurements of attention to data privacy and attention to data security developed using LDA model in the context of GDPR. As robustness checks, we then repeat our analysis using the word count measures.

4. Empirical strategy

We first evaluate our developed measures of data strategies, and particularly firms' attention to data privacy in the context of GDPR. We then explore the determinants of firms' attention to data privacy, and the impact of attention to data privacy on financial performance. We also evaluate the effect of GDPR on attention to data security and third-party data service providers since GDPR also provides guidelines regarding data security and third-party service providers. Our analysis reveals that firms primarily view GDPR as a data privacy regulation and that GDPR primarily affects firms' attention to data privacy.

In the first estimation model that explains firms' attention to data privacy, our primary explanatory variables are exogenous from the point of view of the firm: GDPR regulation and data breaches of other firms in the focal industry. In the second estimation model that explains firms' financial performance (Tobin's Q), our primary explanatory variable, attention to data privacy, is endogenous. We therefore instrument this variable with the GDPR and data breach variables within a System GMM method of estimation.

4.1 Empirical context

A growing number of jurisdictions have adopted new data regulations such as European Union's General Data Protection Regulation (GDPR.eu 2020). The impact of such data regulations on firms' data practices has yet to be investigated. Using our novel measurements, we seek to evaluate the impact of GDPR on firms' attention to data privacy. GDPR is intended to enhance European residents' rights to access the information that firms hold about them and place limitations on what firms can do with their personal data

(GDPR.eu 2020). The regulation was first introduced to the European Parliament in 2012. In March 2014, the European Parliament, the legislative branch of the European Union, adopted GDPR. The European Parliament, the European Council and the European Commission adopted the regulation in its current form in May 2016 followed by a two-year grace period before the full enforcement started on May 25, 2018.

GDPR requires data controllers (i.e., entities that hold personal data and decide how it is to be processed) and data processors (i.e., third-party entities that process personal data on behalf of data controllers) to handle data securely by implementing "appropriate technical and organizational measures." Examples of technical measures include "requiring your employees to use two-factor authentication on accounts where personal data are stored" and "contracting with cloud providers that use end-to-end encryption" (GDPR.eu 2020). Examples of organizational measures include staff training, adding a data privacy policy to employee handbooks, and limiting access to personal data to only those employees in the organization who need it (GDPR.eu 2020). GDPR also mandates implementation of technical measures that enable data controllers to notify data subjects of any breach of their personal data, unless the data controller uses technological safeguards, such as encryption that renders data useless to attackers in the event of a data breach.

In addition, GDPR grants several rights to data subjects including right to be informed, right to access, right to rectification, right to erasure, right to restrict processing, right to data portability, right to object, and rights in relation to automated decision making and profiling. The law requires data controllers to guarantee that these rights of data subjects are respected (GDPR.eu 2020). Article 6 of GDPR explains the instances in which it is legal to process personal data. Examples of such instances include conditions in which data subjects give specific, unambiguous, and informed consent for a specific data processing activity, or within a contract in which data subject is a party. The article prohibits processing of personal data for any other purposes except the ones for which the data subject has provided informed consent (GDPR.eu 2020). Further details on GDPR are provided in Appendix 3.

4.2 Empirical Models and Identification

4.2.1 The Attention to Data Privacy Model

The empirical setting of GDPR provides a suitable context to evaluate the impact of GDPR on firms' attention to data privacy. The empirical setting is particularly suitable for a difference-in-differences (DID) empirical design since GDPR only applies to firms that process European residents' personal data (GDPR.eu 2020). Similar to related studies in the context of GDPR, we exploit the geographic coverage of GDPR (e.g., Aridor et al. 2020, Godinho de Matos and Adjerid 2021, Goldberg et al. 2019, Johnson et al. 2021, Peukert et al. forthcoming) to perform a DID analysis and to identify the effect of the regulation on firms' attention to data privacy. Accordingly, firms with market presence in the E.U. region in 2016 or after, when GDPR was adopted into law, will serve as our treatment group while those firms that do not have market presence in the region during this period will serve as our control group. While prior studies that investigate impacts of data privacy regulations employ the timing of regulatory enforcement as an event study (e.g., Goldfarb and Tucker 2011, Miller and Tucker 2009), we argue that in the context of our research, which measures firms' *attention* to data privacy, the timing of regulatory adoption (i.e., May 25th, 2016) and the start of a two-year grace period before the enforcement date (i.e., May 25th, 2018) is a better proxy for the counterfactual treatment date. Because of the significance of GDPR, we expect that firms increase their attention to practices mandated by the regulation soon after the adoption of GDPR into law. This corresponds to our first hypothesis, H1a.

To address hypothesis H1c, we estimate the effect of data breaches targeting the focal firm's industry peers. Thus, to illuminate the determinants of attention to data privacy, we estimate the following equation:

$$(1)\quad y_{it} = \beta_0 + \beta_1 \text{EU Presence} + \beta_2 \text{EU Presence} \times \text{Post GDPR} + \beta_3 \times \text{Peers' Data Breaches} + \beta_4 \times \text{Controls} + \eta_i + \epsilon_{it}$$

Our primary coefficients of interest are β_2 and β_3. β_2 is an estimate of the average treatment effect under the usual DID identification assumptions (Angrist and Pischke 2008). B_3 is an estimate of the

average treatment effect for peers' data breaches as we assume peers' data breaches to be exogenous to the focal firm. To simplify the notation, t refers to year and i refers to firm. Y_{it} represents our outcomes of interest (attention to data privacy and attention to data security), EU Presence is a dummy that represents whether the company has presence in the European Union and is therefore directly affected by GDPR, and serves as our treatment group. Post GDPR is a dummy variable for the period after the adoption of GDPR into law, and η_i is the firm fixed effect. We also include log(size), log (R&D), and year fixed effects as control variables. To estimate the model, we apply standard fixed-effects methods, because we assume that our key independent variables, GDPR and industry data breaches, are exogenous to the focal firm.

To test for hypothesis H1b, we limit our sample to those firms that have no operations in Europe and thus are not directly exposed to GDPR except through their industry peers. We then estimate the following equation:

$$(2)\quad y_{it} = \beta_0 + \beta_1 \text{Peers' EU Presence} + \beta_2 \text{Peers' EU Presence} \times \text{Post GDPR} + \beta_3 \times \text{Peers' Data Breaches} + \beta_4 \times \text{Controls} + \eta_i + \epsilon_{it}$$

Our primary coefficient of interest in equation (2) is β_2, which is an estimate of the effect of competitive peers' presence in the E.U. after implementation of GDPR on a focal firm's attention to privacy. We operationalize peers' E.U. presence variable as the percentage of firms within the focal firm's industry (defined as 4-digit NAICS code) that have operations in the E.U.. Similar to the model in equation (1), we estimate the model in equation (2) with standard fixed-effects panel methods.

4.2.2 Market Valuation Model

We conduct a Tobin's Q analysis to assess the impact of firms' attention to privacy on performance. Tobin's Q analysis is "the most widely used approach to estimate the value of intangible assets" (Fosfuri and Giarratana 2009, p. 187) such as a firm's data strategies and technological capabilities. Tobin's Q provides a market-based measure of the value of such intangible assets. We estimate an instrumental variable panel model to account for time-variant unobserved effects in ϵ_{it}. Accordingly, we estimate the

following equation to assess the impact of attention to data privacy on financial performance measured as Tobin's Q, corresponding to Hypotheses 2a-c:

$$(3)\quad \text{Tobin}'\text{s Q}_{\text{it}} = \beta_0 + \beta_1 \text{Attention to Privacy}_{it} + \beta_2 \log(ML\ patents)_{it} + \beta_3\ Asia\ Presence_{it} + \beta_4 \log(ML\ patents) \times \text{Attention to Privacy}_{it} + \beta_5 \log(ML\ patents) \times \text{Attention to Privacy}_{it} + \text{Controls} + \eta_{\text{i}} + \epsilon_{\text{it}}$$

We implement a system GMM Instrumental Variable (IV) model (Arellano & Bover, 1995) that is more efficient than 2SLS-type IV methods. It also does not require strict stationarity that the traditional panel 2SLS requires (see e.g. Murtazashvili and Wooldridge, 2008): GMM can accommodate dynamics of the underlying model in the form of first-order autocorrelation. Although the two methods are closely related, GMM is based on less strict assumptions than 2SLS, which is helpful as we do not exactly know the dynamic process that determines firm performance. The GMM estimator produces consistent and efficient coefficient estimates in the presence of endogenous or pre-determined explanatory variables and fixed effects (Blundell and Bond 1998, Roodman 2009).

We use the exogenous variables GDPR and peers' data breaches as instruments to identify the exogenous component of our main endogenous variable, attention to data privacy (hypothesis H2a). We use lagged instances of *the focal firm's number of ML patents* and *peers' presence in Asia* to identify the exogenous components of the focal firm's number of ML patents and Asia presence and to test (H2b and H2c, respectively). We also use past instances of the dependent variable to account for the pre-determined nature of our control variables log(total assets), log(EBITDA ratio) log(CAPX ratio), and log(R&D). We implement the command xtabond2 in the STATA software to estimate the model and report tests for our overidentifying restrictions and second-order autocorrelation.

4.3 Data and Measurement

Attention to Privacy: We constructed a panel of 1,259 publicly traded US firms for which we could retrieve 11,579 10-K reports between 2010 and 2019. Of the 11,579, we were able to automatically parse out 8,707 which we used to develop our measure for attention to data privacy. The HTML code of

the remaining reports was structured differently and would not allow for automatic parsing of the reports into different sections. We followed the procedure explained in section 3 to develop our measures of attention to data privacy which we then merged with the data from Compustat.

Tobin's Q We used Compustat data to construct Tobin's Q as a measure of financial performance. We calculated Tobin's Q = Total Assets – Book Value of Common Equity + Market Value of Common Equity]/Total Assets (Awaysheh et al. 2020). We winsorized the Tobins' Q measure to set the top 0.1 percent of observation equal to the 99.9 percentile of observation because some observations in our sample had significantly higher than average Tobin's Q despite being underperforming firms. Such observations were primarily microcap stocks whose market values significantly exceeded their total assets. Our results are robust to not winsorizing.

Post GDPR: We used a dummy variable for every year after GDPR was adopted into law. GDPR was adopted in May 2016. Depending on a company's revenue, 10-K annual reports must be filed within 60 to 90 days after the company's end of fiscal year (Security Exchange Commission 2009). Since most companies in our sample filed their annual 10-K reports between December of the current year and February of the year after, we assume that most companies had enough time to consider the adoption of GDPR as part of their 2016 10-K filing. Thus, we use a dummy variable for *Post GDPR* for 2016 and after.

EU Presence: We collected data on companies' geographical presence using Compustat geographic segment data. We constructed a dummy variable *EU Presence* which is equal to one if a company had reported any sales in one or more of the E.U. countries in a given year and zero otherwise.

Peers' EU Presence: To test for H1b (i.e., the effect of indirect exposure to GDPR through industry peers), we constructed the *Peers EU Presence* variable as the percentage of a firm's peers that have *EU Presence*. We defined peer firms as firms with the same 4-digit NAICS code as the focal firm after excluding the focal firm itself.

Peers' Data Breaches We use peers' data breaches to identify exogenous variation in firms' attention to data privacy. We collected data on firms' data breaches from Privacy Rights Clearinghouse

(Privacy Rights Clearinghouse 2020). We conducted a fuzzy matching between company names in Privacy Rights Clearinghouse data and company names in Compustat to merge the data breaches data with Compustat data. We constructed the Peers' Data Breaches variable as the number of data breaches experienced by the focal firm's industry peers (defined as firms with the same 4-digit NAICS code as the focal firm after excluding the focal firm itself).

Asia Presence: To test for hypothesis H2c, we constructed a dummy variable called *Asia Presence* for firms' geographical presence in Asia as a proxy for *markets with weak data privacy regimes.* Accordingly, the variable is equal to one if a company had reported any sales in any Asian countries in a given year and zero otherwise.

Peers' Sales in Asia: We constructed a variable for peers' average sales in Asia which we use as an instrumental variable in our GMM model to instrument for a focal firm's presence in Asia. We defined peer firms as firms with the same 4-digit NAICS code as the focal firm after excluding the focal firm itself. Peers' sales in Asia variable captures the average sales of a firm's peers in Asia in millions of US dollars and captures the extent of a firm's peers presence in Asian markets.

Number of machine learning patent applications We used USPTO's classification of machine learning patents (USPTO 2020) to download the data on machine learning patent applications between 2010 and 2019 from Derwent Worlds Patent Index database. We conducted a fuzzy matching between the assignee names in Derwent patent data and company names in Compustat. In total, we identified 11,600 machine learning patent applications by firms in our sample between 2010 and 2019.

Control variables We included log(size) – measured as the number of employees, log (R&D), and year dummies as control variables in equations (1) and (2). We also used log (total assets), EBITDA ratio, CAPX ratio, log(R&D), and year dummies as control variables in equation (3). We used Compustat data to construct these control variables. Table 2 provides the summary statistics for all variables in the analyses. Also, Table 3 presents the pairwise correlations among the variables in the analyses.

4.4. Matching

The baseline difference-in-differences specification in Equation (1) assumes that EU Presence is exogenous and that there are no systematic differences between the first with and without presence in the EU. If that is the case the sample means of all predetermined variables should be the same for firms with and without presence in the EU. To mitigate the potential concern that differences in firm characteristics among E.U.-present firms and non-E.U.-present firms might affect the attention to data strategies, we used Coarsened Exact Matching (CEM) (Iacus et al. 2012, King and Nielsen 2019) to create a matched sample where the observations are balanced by construction. We matched observations with E.U. presence and without E.U. presence on their of *cash holdings*, *earnings per share*, *net income*, *firm size* (i.e., number of employees), and *two-digit NAICS* codes (treated as a categorical variable) between 2010 and 2019 using coarsened exact matching (Iacus et al. 2012). Table 3 presents the summary statistics of E.U.-present and non-E.U.-present observations for both full and matched samples.

*** Insert Table 3 about here ***

5. Results

Our results suggest that the passage of GDPR into law in E.U. member states increased firms' attention to data privacy particularly among the firms that were directly exposed to the regulation due to their presence in the E.U. region. We also provide evidence that firms without a presence in the E.U. increased their attention to data privacy if they were indirectly exposed to the regulation through industry peers. Our findings also show that the increased attention to data privacy affected firms' financial performance. Firms' technological capabilities in digitization as well as their presence in markets with weak privacy regimes moderated the effect on financial performance.

5.1. The Antecedents of Firms' Attention to Data Privacy

We first examine the hypotheses H1a and H1b concerning the drivers of firms' attention to data privacy. To estimate the effect of GDPR on attention to data privacy, we use a DID model as described in Equation (2). The DID estimate assumes a parallel trend in the outcome of study (i.e., attention to data

privacy) between the treatment group (i.e., firms with E.U. presence) and control group (i.e., firms without E.U. presence). Figure 2 provides evidence in support of the parallel pre-trend assumption for the validity of the DID estimator based on the matched sample. The results of the DID estimator in Equation (1) are presented in Table 5. We find evidence in support of an increase in firms' attention to data privacy such that, post-GDPR, firms that have presence in the E.U. economic region on average include extra 1.405 sentences in their 10-K annual reports about data privacy issues. Similarly, using the word count measure, we find evidence in support of an increased attention to data privacy as firms with presence in the E.U. economic region include an additional 1.254 data privacy related words in their 10-K annual reports. These results provide support to our hypothesis H1a.

*** Insert Table 5 about here ***

Next, we focus on understanding the effect of indirect exposure to GDPR through industry peers. For this analysis, we limit our sample to those observations that do not have E.U. presence per the specification in equation (2). We then estimate the interaction term between post GDPR and peers' presence in the E.U. The results of the analysis are provided in Table 6. We find supportive evidence that firms that do not have operations in the E.U. region increased their attention to data privacy as a result of an indirect exposure to the regulation through their industry peers. On average, for every 10% increase in the number of competitive peers that have operations in the E.U., firms included an additional 1.944 sentences about data privacy in their 10-K annual reports. These results are consistent with those we obtain with the word count measure: on average, post GDPR, firms that do not operate in the E.U., included 2.819 additional data privacy-related words in their 10-K annual reports for every 10% increase in their peers' presence in the EU region. Together, these results provide consistent support for hypothesis H1b. Finally, our analysis also provides support for H1c, suggesting that a unit increase in the number of data breaches experienced by competitive peers will on average result in a statistically significant increase of attention to data privacy manifested by an additional 0.0102 sentences about data privacy included in the focal firm's 10-K annual report. Similarly, using our word count measure, we find that a unit increase

in competitive peers' data breaches will result in an additional 0.0126 data privacy-related words included in the focal firm's 10-K annual report.

*** Insert Table 6 about here ***

5.1. The Effect of Attention to Data Privacy on Performance

Next, we turn on our attention to testing hypotheses H2a – H2c. We apply dynamic GMM (generalized method of moments) estimators (Arellano and Bover 1995, Blundell and Bond 1998), which incorporate the dynamic nature of the financial performance variable Tobin's Q and utilize instruments that control for time-variant unobserved heterogeneity and simultaneity (Wintoki et al. 2012). We use Stata xtabond2 instruction in our estimation (Roodman 2009). We limit the number of instruments by using the 'collapse' function as in Roodman (2009), which creates one instrument for each variable and lag distance, rather than one for each time period, variable and lag distance. The collapse option significantly increases the power of the Sargan test of over-identification. Our reported specification tests here include second-order autocorrelation and Sargan test statistic (and its p-value).

We use the lagged dependent variable, GDPR, and peers' data breaches as instruments to identify the exogenous variation in firms' attention to privacy. We also use peers' sales in Asia to instrument for firms' presence in Asia, which is a proxy for access to markets with weak data privacy regimes. We instrument ML patenting with lagged instances of industry peers' ML patenting. The baseline model for H2a is presented in model (1) of Table 7.

We find supportive evidence that increased attention to data privacy is associated with a decrease in financial performance. The increased attention to data privacy manifested as an extra sentence of data privacy content in a 10-K report is, on average, associated with decrease of 4.919 in Tobin's Q. This result supports hypothesis H2a.

*** Insert Table 7 about here ***

Next, we find that the attention to data privacy and machine learning innovation are complementary in their effects on firms' performance. According to model (3) in Table 7, if firms that

need to respond to the regulatory pressure on their data privacy practices also are able to generate new patentable machine-learning technologies, they can mitigate the adverse effects of the regulation and even benefit from it if they are highly successful in their invention activities. Accordingly, controlling for the extent of attention data privacy, an additional ML patent filed by the focal firm is on average associated with 3.185% increase in the firm's Tobin's Q which supports hypothesis H2b. We also find that attention to data privacy and presence in markets with weak data privacy regimes are complementary in their effects on firms' performance. According to models (5) and (6) in Table 7, we find that if firms that are under regulatory pressure for their data privacy practices are also present in markets with weak data privacy regulations, they can mitigate the adverse effects of the exposure to a stringent data privacy regulation such as GDPR and even benefit from it if they are able to substitute the data sources required for their innovation activities from regions where data privacy regimes are weak. The interdependence between the attention to data privacy and presence in markets with weak data privacy regimes is demonstrated in models (5) and (6) of Table 7. Accordingly, controlling for the extent of attention data privacy, a focal firm's presence in Asian markets is on average associated with 28.15 increase in the firm's Tobin's Q which provides evidence in support of H2c.

*** Insert Table 7 about here ***

4. Supplementary Analyses

Impact of GDPR on technological capabilities in digitization: We proposed that firms' market presence in weak data privacy regimes may have a moderating effect on firms' performance by allowing firms to build capabilities in data-driven innovation (e.g., machine learning) while incurring lower costs of innovation. If our proposed mechanism holds, we should see an increased level of innovation in technological areas that rely heavily on personal data. Thus, controlling for attention to data privacy, we would expect firms with presence in markets with weak data privacy regimes to more successfully innovate in areas related to machine learning. We apply dynamic GMM estimators to test the impact of presence in markets with data privacy regimes post GDPR. Our instruments again include the lagged

dependent variable, GDPR, and peers' data breaches. This analysis suggests that after controlling for firm's attention to data privacy, firms that are present in markets with weak data privacy regimes are more likely to innovate in areas that require access to personal data compared to firms that do not have such presence. These results are presented in Table 8.

*** Insert Table 8 about here ***

Alternative measure of technological capabilities in digitization: As a supplementary analysis, we developed an alternative measure of technological capabilities in digitization by counting the number of patents filed by firms in our sample that either have words "data" and "privacy" or data" and "security" in their abstracts. We used Derwent Worlds Patent Index database for this purpose. We conducted a fuzzy matching between the assignee names in Derwent patent data and company names in Compustat. In total, we identified 3,838 patents filed by firms in our sample that met the search criteria. We repeated our test of hypotheses H2b using the alternative measure and found consistent evidence for the moderating effect of technological capabilities in digitization on firm's performance. The results of this analysis are presented in Table 9.

*** Insert Table 9 about here ***

Impact of Attention to Data Privacy on Data Breaches: In this paper, we argued that firm's attention to data privacy is costly and that it impacts firm's performance. One may argue that our approach to measuring firm's attention to data privacy does not capture the firms' substantive attention to data privacy and that 10-K reports are primarily legal documents that are disconnected from firms' substantive actions. To evaluate this argument, we investigate the impact of increased attention to data privacy as a result of GDPR and peers' data breaches. Using the same set of instruments, we apply our GMM model to evaluate the impact of attention to data privacy on the focal firm's data breaches. Our analysis provides evidence of reduced data breaches as a result of increased attention to data privacy caused, among other things, by the exogenous drivers GDPR and peers' data breaches. These results are presented in Table 10..

*** Insert Table 10 about here ***

Placebo treatment year: We ran placebo tests to assess the potential for false positive effects. We first chose a counterfactual treatment date ranging from year 2011 to 2018 which includes the actual treatment year of 2016. We then constructed a subsample using a two-year window around the treatment year (e.g., for placebo treatment year equal to 2015, we constructed a subsample of our panel from 2014 to 2016). We then estimated β_2 the equation (1) for the constructed subsample. Figure 7 illustrates these results for different counterfactual treatment years. The estimates for placebo treatment years 2011, 2014, and 2015 are not statistically significant from zero. The estimates for placebo treatment years 2012 and 2013 are statistically different from zero; however, they are much smaller than the actual treatment estimate for 2016. Note that the year 2012 coincides with the European Commission's proposal to strengthen online privacy rights and digital economy which was the very first attempt that laid the foundations for GDPR (GDPR.eu 2020). Thus, the non-zero estimated effect for years 2012 and 2013 may attributed to E.U. firms' attention to the original proposal.

*** Insert Figure 7 about here ***

Placebo dependent variable our final concern is that other factors such as technological trends may be influencing firms' attention to data practices including data privacy. To rule out this concern, we examine other aspects of firms' data practices as placebo dependent variables. We already investigated the effect of GDPR on firms' attention to data privacy and data security in Table 5. We now test for the effect of GDPR on firms' attention to other data practices. Given the scope of GDPR, we do not expect changes in firms' attention to other data practices. The results are reported in Table 11. As expected, there are no statistically significant relationships between GDPR and most other data practices, except a small and marginally significant increase ($p < 0.1$) in firms' attention to third-party data service providers. This is understandable, given that GDPR regulates data processing arrangements between focal firms and third-party organizations (GDPR.eu 2020). We also find no effect of peers' data breaches on firms'

attention to other data practices. Overall, we feel confident that we are not simply capturing a secular trend to increase attention to data practices.

*** Insert Table 11 about here ***

5. Discussion and conclusions

In this study we present a topic modeling approach for discovering and analyzing the value of firms' attention to data privacy across four diverse sectors: Information, finance, healthcare, and retail. Our first contribution is to demonstrate that we can formulate informative variables about firms' attention to data-related practices, such as data privacy, from annual report disclosures. Our findings suggest that firms have distinct data practices that evolve over time and that vary meaningfully across firms and across industries. In particular, firms' attention to data privacy responds to exogenous changes in regulation, competition, and industry norms and expectations. Firms that have a commercial presence in the E.U. strongly increase their attention to data privacy after the adoption of the GDPR. This adoption effect spills over to competitive peers, as firms with many peers that have E.U. presence also significantly increase their attention to data privacy.

Our second contribution highlights that firms' attention to data privacy has substantial performance effects. The direct effect of increasing attention to data privacy is negative: Tobin's Q decreases when firms grow their attention to data privacy. This suggests that, generally, firms are not able to recoup their investments in data privacy from their customers in the form of stronger demand. We speculate that this is because it is difficult for customers to distinguish firms that have made such valuable investments on their behalf. Thus, although customers tend to appreciate privacy goods, firms are unable to credibly communicate about these investments, and customers are not able to compensate firms for the privacy goods provided.

Our third contribution shows that firms have heterogeneous capabilities to develop their practices related to data privacy. We test for two moderating effects: technological capabilities related to data analytics (machine learning) and access to alternative sources of personal data. We find that both

variables positively moderate the performance impact of attention to data privacy. These results imply that firms with technological capabilities and access to data in jurisdictions with weak regulation of data privacy can alleviate the costs and negative revenue implications of data privacy practices. Moreover, the moderating effects suggest that practices related to data privacy are complemented by other assets and practices. Data privacy is an expensive service feature to provide for customers who may not even notice it until it fails. Privacy features also tend to reduce the firm's access to data for product and service innovation. Post-GDPR, E.U. markets for external data about consumers have significantly dried up. As a result, firms must make the most of the customer data they continue to access. Machine learning patents indicate firms that have strong analytics capabilities in-house. Thus, in a post-GDPR world, market power has shifted toward firms with strong proprietary technological assets related to data privacy. Alternatively, firms can seek data from other, less regulated jurisdictions. Thus, firms with presence in Asian economies face lesser performance penalties from increased attention to data privacy.

If firms are indeed unable to differentiate themselves through enhanced attention to privacy in the marketplace because consumers either are unable to discern higher versus lower quality data privacy practices or are not willing to pay much for the former, then firms will not voluntarily fix privacy issues related to their users' data. In that case, the only remedies to address the significant societal and consumer harms are through regulations such as GDPR that raise all firms' costs. While these results are novel in the context of data privacy, we can relate them to the context of environmental harms such as pollution or carbon emissions. It is generally difficult for firms to profit from investments in Corporate Social Responsibility investments. If a polluting firm voluntarily reduces its emissions, it adds costs without necessarily being able to recoup them in the form of higher prices enabled by differentiation. However, in case regulation is imminent in the future, firms may benefit from investing early in regulatory compliance of processes and practices. Data privacy can thus be viewed as an issue with negative externalities that need to be addressed via regulation rather than relying on firms' voluntary efforts to offer privacy goods to their customers.

References

Abbasi A, Sarker S, Chiang R (2016) Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *JAIS* 17(2):1–32.

Adner R, Puranam P, Zhu F (2018) Special Issue of Strategy Science: Strategy in the Digital Era. *Strategy Science*.

Angrist JD, Pischke JS (2008) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).

Aragòn-Correa JA, Marcus AA, Vogel D (2019) The Effects of Mandatory and Voluntary Regulatory Pressures on Firms' Environmental Strategies: A Review and Recommendations for Future Research. *ANNALS* 14(1):339–365.

Arellano M, Bover O (1995) Another look at the instrumental variable estimation of error-components models. *Journal of econometrics* 68(1):29–51.

Aridor G, Che YK, Nelson W, Salz T (2020) The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR. *SSRN Journal*.

Awaysheh A, Heron RA, Perry T, Wilson JI (2020) On the relation between corporate social responsibility and financial performance. *Strategic Management J.* 41(6):965–987.

Bao Y, Datta A (2014) Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures. *Management Science* 60(6):1371–1391.

Barrett L (2018) Confiding in Con Men: US Privacy Law, the GDPR, and Information Fiduciaries. *Seattle UL Rev.* 42:1057.

Benaroch M, Chernobai A (2017) Operational IT failures, IT value-destruction, and board-level IT governance changes. *MIS Quarterly, Forthcoming*.

Bessen JE, Impink SM, Reichensperger L, Seamans R (2020) *GDPR and the Importance of Data to AI Startups* (Social Science Research Network, Rochester, NY).

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. *Journal of machine Learning research* 3(1):993–1022.

Blundell R, Bond S (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of econometrics* 87(1):115–143.

Borgman CL (2012) The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63(6):1059–1078.

Brynjolfsson E, Hitt LM, Yang S (2002) Intangible assets: Computers and organizational capital. *Brookings papers on economic activity* 2002(1):137–181.

Brynjolfsson E, McElheran K (2016) The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review* 106(5):133–139.

Bundy J, Pfarrer MD, Short CE, Coombs WT (2017) Crises and crisis management: Integration, interpretation, and research development. *Journal of management* 43(6):1661–1692.

Chebli O, Goodridge PR, Haskel J (2015) Measuring activity in big data: New estimates of big data employment in the UK market sector.

Chen, Chiang, Storey (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36(4):1165.

Corritore M, Goldberg A, Srivastava SB (2019) Duality in Diversity: How Intrapersonal and Interpersonal Cultural Heterogeneity Relate to Firm Performance. *Administrative Science Quarterly*:000183921984417.

De Finetti B (1990) Theory of Probability, Vol. I (1974). *John Wiley & Sons* 5(8):17.

Eggers JP, Kaplan S (2009) Cognition and Renewal: Comparing CEO and Organizational Effects on Incumbent Adaptation to Technical Change. *Organization Science* 20(2):461–477.
Ferrara P, Pistoia M, Tripp O (2020) Privacy detection of a mobile application program.
Fosfuri A, Giarratana MS (2009) Masters of War: Rivals' Product Innovation and New Advertising in Mature Product Markets. *Management Science* 55(2):181–191.
Fremeth AR, Shaver JM (2014) Strategic Rationale for Responding to Extra-Jurisdictional Regulation: Evidence from Firm Adoption of Renewable Power in the Us. *Strategic Management Journal* 35(5):629–651.
Gamache DL, McNamara G, Mannor MJ, Johnson RE (2014) Motivated to Acquire? The Impact of CEO Regulatory Focus on Firm Acquisitions. *AMJ* 58(4):1261–1282.
GDPR.eu (2020) General Data Protection Regulation (GDPR) Compliance Guidelines. *GDPR.eu*. Retrieved (September 30, 2020), https://gdpr.eu/.
Godinho de Matos M, Adjerid I (2021) Consumer Consent and Firm Targeting After GDPR: The Case of a Large Telecom Provider. *Management Science*.
Goldberg S, Johnson G, Shriver S (2019) Regulating Privacy Online: An Economic Evaluation of the GDPR. *SSRN Journal*.
Goldfarb A, Tucker CE (2011) Privacy regulation and online advertising. *Management science* 57(1):57–71.
Hall RE (2001) The stock market and capital accumulation. *American Economic Review* 91(5):1185–1202.
Hannigan T, Haans R, Vakili K, Tchalian H, Glaser V, Wang M, Kaplan S, D. Jennings D (2019) Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*.
Hoberg G, Phillips G (2010) Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *The Review of Financial Studies* 23(10):3773–3811.
Iacus SM, King G, Porro G (2012) Causal Inference Without Balance Checking: Coarsened Exact Matching. *Political analysis*:1–24.
Johnson G, Shriver S, Goldberg S (2021) *Privacy & Market Concentration: Intended & Unintended Consequences of the GDPR* (Social Science Research Network, Rochester, NY).
Khan LM (2019) The separation of platforms and commerce. *Columbia Law Review* 119(4):973–1098.
King G, Nielsen RA (2019) Why propensity scores should not be used for matching.
Kitchin R (2014) *The data revolution: Big data, open data, data infrastructures and their consequences* (Sage).
Koutroumpis P, Leiponen A, Thomas LD (2020) Small is big in ICT: The impact of R&D on productivity. *Telecommunications Policy* 44(1):101833.
Koutroumpis P, Leiponen A, Thomas LDW (2020) Markets for data. *Industrial and Corporate Change* 29(3):645–660.
Leiblein MJ, Reuer JJ, Zenger T (2018) What makes a decision strategic? *Strategy Science* 3(4):558–573.
Lycett M (2013) 'Datafication': making sense of (big) data in a complex world. *European Journal of Information Systems* 22(4):381–386.

Martin K (2020) Breaking the Privacy Paradox: The Value of Privacy and Associated Duty of Firms. *Bus. Ethics Q.* 30(1):65–96.
McAfee Labs (2016) *McAfee Labs Threat Reports.* (Santa Clara, CA).
Miller AR, Tucker C (2009) Privacy protection and technology diffusion: The case of electronic medical records. *Management science* 55(7):1077–1093.
National Science Foundation (2019) NSF's 10 Big Ideas - Harnessing the Data Revolution. Retrieved (February 15, 2020), https://www.nsf.gov/news/special_reports/big_ideas/harnessing.jsp.
Ocasio W (1997) Towards an attention-based view of the firm. *Strategic management journal* 18(S1):187–206.
Ocasio W, Joseph J (2018) The Attention-Based View of Great Strategies. *Strategy Science* 3(1):7.
Peukert C, Bechtold S, Batikas M, Kretschmer T (forthcoming) Regulatory Spillovers and Data Governance: Evidence from the GDPR. *Marketing Science*.
Privacy Rights Clearinghouse (2020) Data Breaches | Privacy Rights Clearinghouse. Retrieved (September 30, 2020), https://privacyrights.org/data-breaches.
Roodman D (2009) How to do xtabond2: An introduction to difference and system GMM in Stata. *The stata journal* 9(1):86–136.
Saunders A, Tambe P (2015) Data Assets and Industry Competition: Evidence from 10-K Filings. *Available at SSRN 2537089*.
Security Exchange Commission (2005) SEC regulation S-K, item 503(c). https://www.sec.gov/Archives/edgar/data/60086/000119312505057094/filename3.htm.
Security Exchange Commission (2009) SEC.gov \textbar Form 10-K. https://www.sec.gov/fast-answers/answers-form10khtm.html.
Tambe P (2014) Big data investment, skills, and firm value. *Management Science* 60(6):1452–1469.
Tambe P, Hitt LM (2012) The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research* 23(3-part-1):599–617.
Tang J, Meng Z, Nguyen X, Mei Q, Zhang M (2014) Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis. *International Conference on Machine Learning*. 190–198.
Thomas L, Leiponen A, Koutroumpis P (2021) Profiting from data: Business models for data products. Cennamo C, GB D, F Z, eds. *Elgar Handbook of Research on Digital Strategy*.
Uhlir PF, Cohen D (2011) Internal Document. Board on Research Data and Information, Policy and Global Affairs Division. *National Academy of Sciences* 18.
USPTO (2020) *Inventing AI: Tracing the diffusion of artificial intelligence with U.S. patents*
Vasudeva G, Nachum L, Say GD (2018) A signaling theory of institutional activism: How Norway's sovereign wealth fund investments affect firms' foreign acquisitions. *Academy of Management Journal* 61(4):1583–1611.
Verizon (2017) *Data breach investigations report* (Verizon, New York).
Villalonga B (2004) Intangible resources, Tobin'sq, and sustainability of performance differences. *Journal of Economic Behavior & Organization* 54(2):205–230.
Wintoki MB, Linck JS, Netter JM (2012) Endogeneity and the dynamics of internal corporate governance. *Journal of financial economics* 105(3):581–606.
Wu L, Hitt L, Lou B (2020) Data analytics, innovation, and firm productivity. *Management Science* 66(5):2017–2039.

York JG, Vedula S, Lenox MJ (2018) It's Not Easy Building Green: The Impact of Public Policy, Private Actors, and Regional Logics on Voluntary Standards Adoption. *AMJ* 61(4):1492–1523.

Figure 1: Overview of the Measurement Development Approach Using LDA

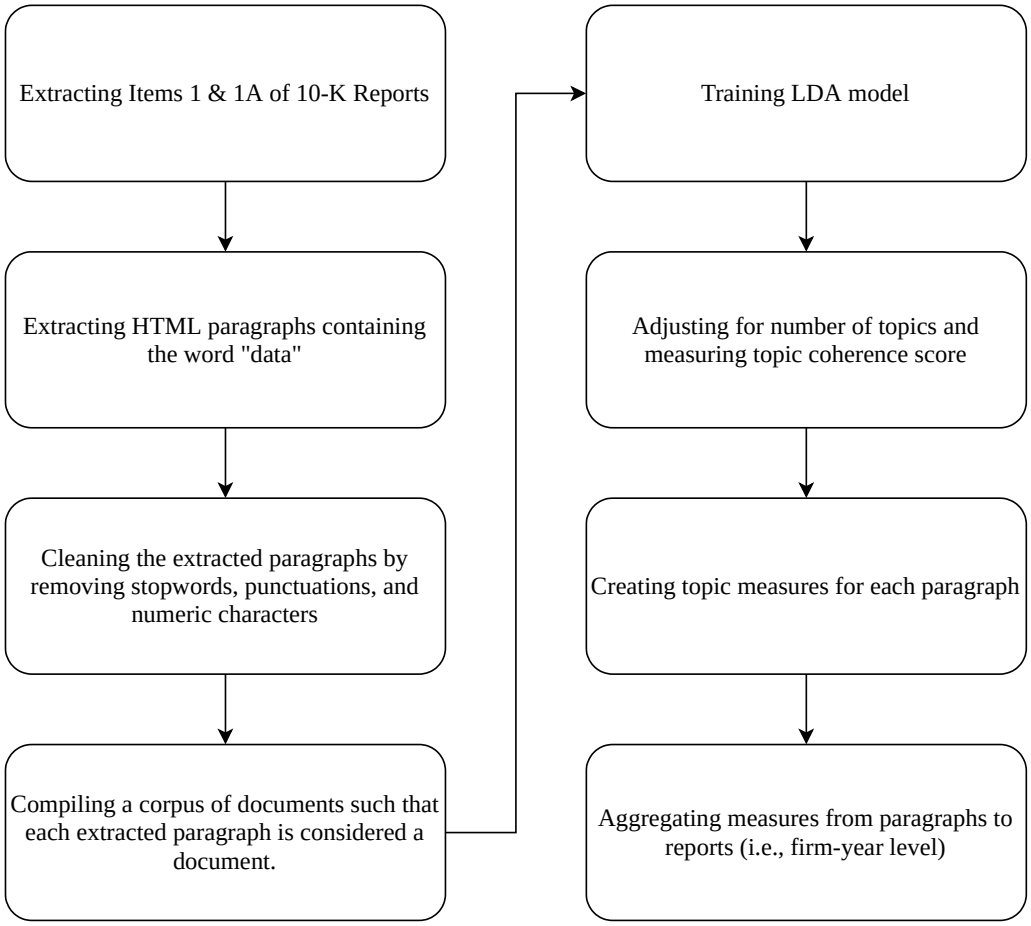


Figure 2: Attention to Data Privacy for Firms with and without Operations in the E.U.

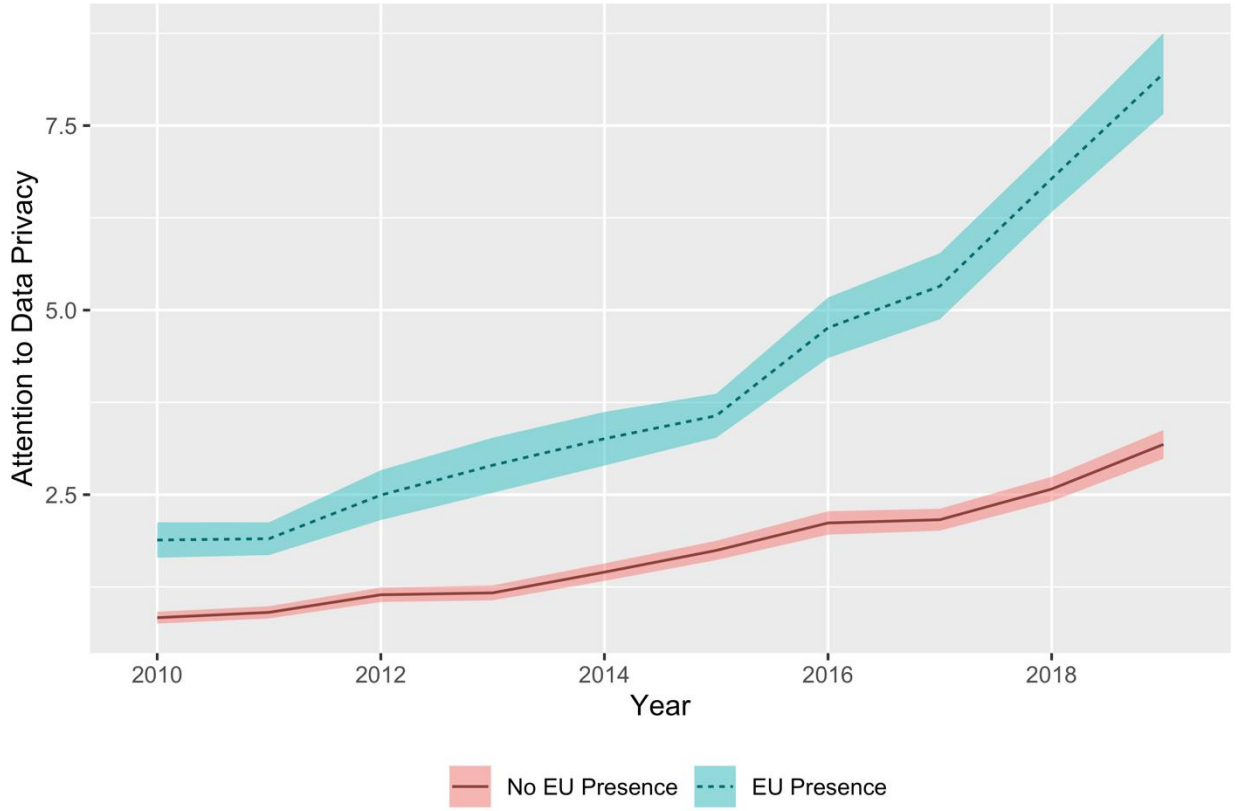


Table 1: Topic Assignments

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
Data Security	**Data Systems Failure**	**Data Infrastructure**	**Data-driven Products & Services**	**Data Privacy**	**Data Management Systems**	**Third-Party Data Services**
data	data	data	service	data	data	party
security	result	center	data	law	database	third
information	system	customer	product	regulation	application	data
customer	loss	data_center	market	privacy	software	third_party
breach	business	business	company	protection	technology	provider
data_security	operation	service	investment	state	market	service
system	failure	ability	certain	data_protection	based	service_provider
security_breach	information	system	customer	requirement	management	processing
access	damage	facility	product_service	information	platform	system
employee	attack	infrastructure	insurance	security	user	rely

Table 2: Summary Statistics

	Mean	SD	Min	Max	N
Tobin's Q	2.06	5.61	0.34	141.74	8123
log (size)	1.20	1.20	0.00	6.03	8290
log (Cash holdings)	4.41	1.85	0.00	10.40	8290
log (COGS)	4.98	2.39	-2.02	11.99	8277
log (revenue)	6.07	2.04	0.00	12.18	8288
log (operating profit)	5.38	1.87	-3.91	11.35	8124
log (total assets)	7.55	2.07	0.01	14.61	8290
EBITDA Ratio	-0.14	12.70	-867.80	2.41	8091
CAPEX Ratio	0.05	0.42	-0.00	33.86	8208
log (R&D)	0.69	1.60	0.00	9.48	8290
Post GDPR	0.40	0.49	0.00	1.00	8290
EU Presence	0.24	0.42	0.00	1.00	8290
Firm's Number of Breaches	0.03	0.33	0.00	16.00	8290
Peers' Data Breaches	13.23	12.92	0.00	41.00	8290
Attention to Privacy (topic modeling)	2.37	3.93	0.00	38.77	5977
Attention to Security (topic modeling)	1.91	2.39	0.00	28.27	5977
Attention to Third parties (topic modeling)	1.37	1.98	0.00	23.19	5977
Attention to Privacy (word count)	2.29	4.48	0.00	54.00	5977
Attention to Security (words count)	6.89	9.65	0.00	172.00	5977
log (ML patents)	0.05	0.32	0.00	5.49	8290
log (Privacy patents)	0.04	0.28	0.00	3.85	8290
Asia Presence	0.02	0.15	0.00	1.00	8290
% Peers EU Presence	2.38	2.18	0.00	16.67	8290

Table 3: Correlation Matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)
(1) Tobin's Q	1.00																						
(2) log (Size)	-0.01	1.00																					
(3) log (cash holding)	-0.00	0.57	1.00																				
(4) log (COGS)	-0.03	0.81	0.62	1.00																			
(5) log (revenue)	-0.06	0.86	0.74	0.94	1.00																		
(6) log (operating profit)	-0.05	0.80	0.75	0.77	0.92	1.00																	
(7) log (total assets)	-0.24	0.55	0.71	0.56	0.73	0.77	1.00																
(8) EBITDA Ratio	-0.15	0.02	0.02	0.05	0.07	0.11	0.06	1.00															
(9) CAPEX Ratio	0.01	-0.02	-0.04	-0.04	-0.06	-0.08	-0.06	-0.08	1.00														
(10) log (R&D)	0.27	-0.01	0.16	-0.02	0.02	0.11	-0.18	-0.01	0.01	1.00													
(11) Post GDPR	0.01	0.01	0.01	0.03	0.04	0.05	0.05	0.00	-0.01	0.05	1.00												
(12) EU Presence	0.11	0.15	0.26	0.20	0.20	0.20	0.00	0.01	-0.01	0.38	0.01	1.00											
(13) Firm's Number of Breaches	-0.00	0.13	0.13	0.13	0.14	0.13	0.11	0.00	-0.00	-0.01	-0.01	0.02	1.00										
(14) Peers' Data Breaches	-0.13	-0.20	0.04	-0.18	-0.12	-0.09	0.22	0.02	-0.02	-0.24	-0.08	-0.13	0.06	1.00									
(15) Attention to Privacy	0.19	0.07	0.16	0.10	0.11	0.14	-0.09	-0.02	0.00	0.48	0.20	0.27	0.06	-0.19	1.00								
(16) Attention to Security	0.20	0.12	0.16	0.14	0.14	0.17	-0.05	-0.02	-0.00	0.41	0.23	0.21	0.05	-0.19	0.69	1.00							
(17) Attention to Privacy (word count)	0.17	0.06	0.15	0.09	0.10	0.13	-0.10	-0.01	0.00	0.48	0.21	0.26	0.07	-0.19	0.89	0.65	1.00						
(18) Attention to Security (words count)	0.16	0.13	0.17	0.14	0.16	0.19	0.03	-0.02	-0.01	0.34	0.26	0.18	0.07	-0.14	0.63	0.84	0.63	1.00					
(19) Attention to Third-parties	0.17	-0.03	0.09	0.02	0.02	0.05	-0.10	0.00	0.02	0.44	0.15	0.18	0.03	-0.14	0.62	0.63	0.53	0.47	1.00				
(20) log (ML patents)	0.14	0.11	0.17	0.09	0.13	0.17	0.07	0.00	0.00	0.40	0.06	0.16	0.05	-0.06	0.24	0.16	0.23	0.17	0.11	1.00			
(21) log (Privacy patents)	0.12	0.10	0.21	0.10	0.14	0.18	0.09	0.00	-0.00	0.31	-0.03	0.15	0.07	-0.03	0.20	0.20	0.17	0.23	0.11	0.50	1.00		
(22) Asia Presence	0.05	0.07	0.08	0.05	0.07	0.09	0.02	0.00	-0.01	0.15	0.00	0.22	0.00	-0.03	0.03	0.04	0.02	0.03	0.04	0.02	0.02	1.00	
(23) Peers EU Presence	0.29	0.09	0.15	0.17	0.11	0.08	-0.27	-0.04	0.02	0.59	0.11	0.49	0.00	-0.27	0.50	0.45	0.49	0.36	0.41	0.21	0.21	0.11	1.00

Table 4: Mean Differences Between Treatment and Control Groups Before and After Matching

	Before Matching			After Matching		
	Means Treated	**Means Control**	**Std. Mean Diff.**	**Means Treated**	**Means Control**	**Std. Mean Diff.**
Distance	0.34	0.20	0.78	0.32	0.31	0.04
Size (thousand employees)	30.00	8.40	0.15	13.88	10.66	0.02
EPS (Dollars)	2.00	1.60	0.06	1.60	1.28	0.06
Net income (Millions)	1000.00	300.00	0.21	520.40	372.40	0.04
Cash holdings (Millions)	2100.00	550.00	0.16	811.80	477.30	0.03
NAICS 44	0.05	0.07	-0.10	0.05	0.05	0.00
NAICS 45	0.04	0.03	0.07	0.03	0.03	0.00
NAICS 51	0.55	0.20	0.70	0.56	0.56	0.00
NAICS 51	0.32	0.64	-0.67	0.32	0.32	0.00
NAICS 62	0.03	0.05	-0.15	0.03	0.03	0.00

Table 5: Fixed Effects Estimations of Attention to Data Privacy

	(1) **Attention to privacy (Topic modeling)**	(2) **Attention to privacy (Word count)**
log(size)	0.857***	0.796***
	(0.135)	(0.161)
log(R&D)	0.627***	0.798***
	(0.0824)	(0.0983)
Post GDPR	1.716***	2.360***
	(0.124)	(0.149)
EU presence	-0.469***	-0.539***
	(0.124)	(0.148)
EU Presence × Post GDPR	1.405***	1.254***
	(0.101)	(0.121)
Peers' Data Breaches	0.0102**	0.0126**
	(0.00326)	(0.00389)
Constant	0.0401	-0.287
	(0.187)	(0.223)
N	5970	5970
Adjusted R^2	0.080	0.070
Firm fixed effects	Yes	Yes
Year fixed effects	Yes	Yes

Standard errors in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Fixed Effects Estimations of Competitive Peers' Presence in EU & Attention to Privacy

	(1) **Attention to Privacy (Topic modeling measure)**	(2) **Attention to Privacy (Word count measure)**
log (size)	0.806***	0.945***
	(0.133)	(0.154)
log (R&D)	0.663***	0.933***
	(0.101)	(0.118)
Post GDPR	1.119***	1.351***
	(0.122)	(0.142)
Peers' data breaches	0.0104***	0.00947**
	(0.00307)	(0.00357)
Peers EU presence	0.022	0.024
	(4.613)	(5.370)
Peers EU Presence × Post GDPR	0.194***	0.282***
	(0.023)	(0.027)
Constant	-0.166	-0.660**
	(0.185)	(0.215)
N	4487	4487
Adjusted R^2	0.015	0.067
Firm fixed effects	Yes	Yes
Year fixed effects	Yes	Yes

Standard errors in parentheses
*+ p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001*

Table 7: GMM IV Estimation of Tobin's Q

	(1) Tobin's Q	(2) Tobin's Q	(3) Tobin's Q	(4) Tobin's Q	(5) Tobin's Q	(6) Tobin's Q
log (total assets)	11.93*	11.39*	16.05**	6.910***	6.042*	8.050*
	(4.838)	(5.757)	(5.343)	(2.056)	(2.369)	(3.200)
EBITDA ratio	-0.210*	-0.290+	-0.207+	-0.283**	-0.275+	-0.306*
	(0.0960)	(0.154)	(0.125)	(0.105)	(0.159)	(0.155)
CAPX ratio	0.130	0.896	5.721	-4.551	-3.360	4.446
	(3.818)	(5.498)	(4.527)	(3.222)	(4.122)	(5.469)
log (R&D)	16.19***	9.619**	10.01***	11.10**	10.98*	12.75*
	(1.969)	(3.426)	(2.624)	(4.267)	(5.118)	(6.117)
Attention to privacy	-4.919***	-1.855*	-2.376***	-4.291***	-3.655**	-3.210*
	(0.741)	(0.728)	(0.631)	(1.063)	(1.334)	(1.385)
log (ML patents)		-3.318+	-33.53**			-27.30
		(1.825)	(10.38)			(21.17)
log (ML patents) × Attention to privacy			4.210**			3.191
			(1.477)			(2.630)
Asia presence				55.48	-131.7	-165.1+
				(44.25)	(84.94)	(89.04)
Asia presence × Attention to privacy					25.02**	22.93*
					(9.392)	(8.936)
Constant	-85.22**	-35.43	-89.47**	-50.17**	-52.96**	-59.17*
	(30.98)	(24.21)	(32.00)	(15.48)	(17.97)	(24.34)
N	5803	5803	5803	5803	5803	5803
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
AR (2)	0.519	-0.411	-1.547	-1.732	-1.912	-1.948
AR (2) (p)	0.604	0.681	0.122	0.083	0.056	0.051
Sargan test	4.435	11.469	8.582	4.397	3.664	1.440
Sargant test (p)	0.350	0.177	0.284	0.355	0.994	0.998

Standard errors in parentheses. One-step system GMM. We formed Roodman-collapsed GMM-instruments from the lagged dependent variable, GDPR, peers' data breaches, and peers' sales in Asia.

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix 1: Labeled Topics and Representative Examples

Topic	Representative Example
Topic 1 – Data Security	Although we have developed systems and processes that are designed to protect our data and customer data and to prevent data loss and other security breaches and expect to expend significant additional resources to bolster these protections these security measures cannot provide absolute security our information technology and infrastructure may be vulnerable to cyberattacks or security breaches and third parties may be able to access our customers' personal or proprietary information and payment card data that are stored on or accessible through those systems. (Paypal, 2016)
Topic 2 – Data Systems Failure	Network and information systems-related events such as computer hackings cyber-attacks ransomware computer viruses worms or other destructive or disruptive software process breakdowns denial of service attacks malicious social engineering or other malicious activities or any combination of the foregoing or power outages natural disasters terrorist attacks or other similar events could result in damage to our property equipment and data affect our ability to maintain ongoing operations and result in significant expenditures to repair or replace the damaged property or information systems reacquire access to networks and information systems or to protect them from similar events in the future. (Royal Gold Inc., 2018)
Topic 3 – Data Infrastructure	Oracle Cloud Services are designed to provide comprehensive software and hardware management and maintenance services for customers hosted at Oracle data center facilities, select partner data centers or physically on-site at customer facilities. Advanced Customer Services provides support services, both onsite and remote, to Oracle customers to enable increased performance and higher availability of their products and services. (Oracle, 2011)
Topic 4 – Data-driven products and services	Demand for smartphones and data services continues to grow across all of our wireless markets and our value to our customers in some markets depends in part on our network's ability to provide high-quality and high-capacity network service to smartphone devices. (Atlantic Telenetwork, Inc, 2017)

Topic 5 – Data privacy	For example, the EU adopted a comprehensive General Data Protection Regulation (the "GDPR"), which came into effect in May 2018, as supplemented by any national laws (such as in the U.K., the Data Protection Act 2018) and further implemented through binding guidance from the European Data Protection Board, and expanded the scope of the EU data protection law to foreign companies processing personal data of European Economic Area ("EEA") individuals, imposed a stricter data protection compliance regime, and included new data subject rights (e.g., the right to erasure, commonly known as the "right to be forgotten"). (Paypal, 2019)
Topic 6 – Data management systems	Our standardized platform includes the most comprehensive proprietary database in the industry; the largest research department in the industry; proprietary data collection information management and quality control systems; a large in-house product development team; a broad suite of web-based information analytics and marketing services; a large team of analysts and economists; and a large base of clients. (CoStar Group, 2014)
Topic 7 – Third party data services	We may have to dedicate significant resources to manage risks and regulatory burdens presented by our relationship with vendors and third-party service providers including our data processing and cybersecurity service providers. (C&F financial corporation, 2016)

Appendix 2: Example of Attention to Data Strategies Across Different Industries

Figure 3: Attention to Data Strategies – Facebook

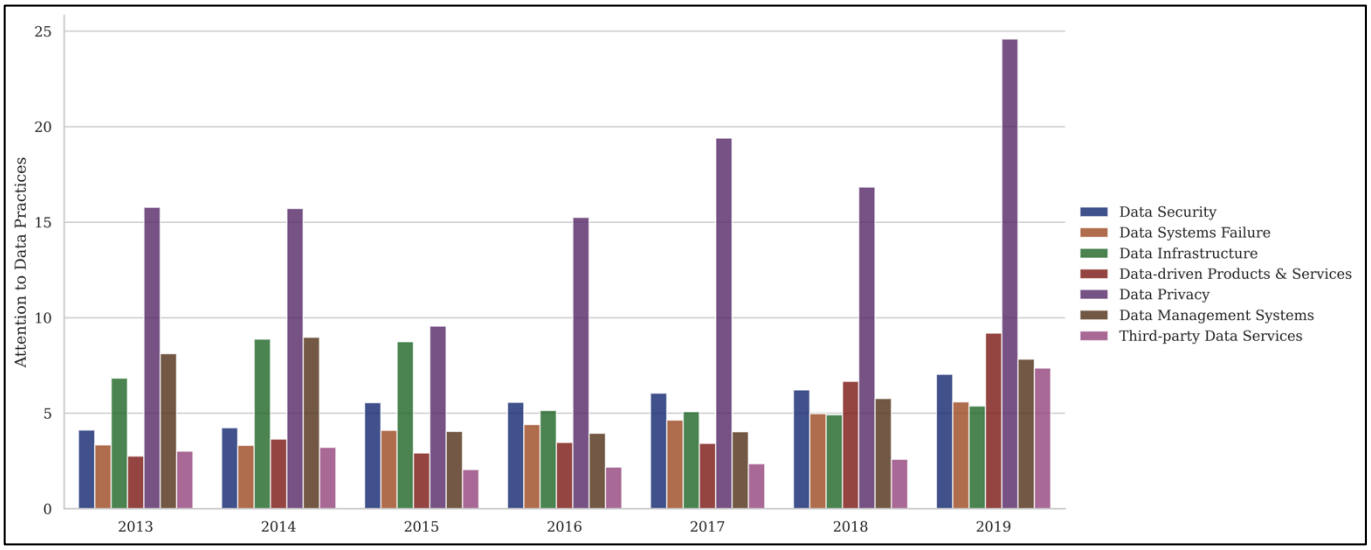


Figure 4: Attention to Data Strategies – Amazon

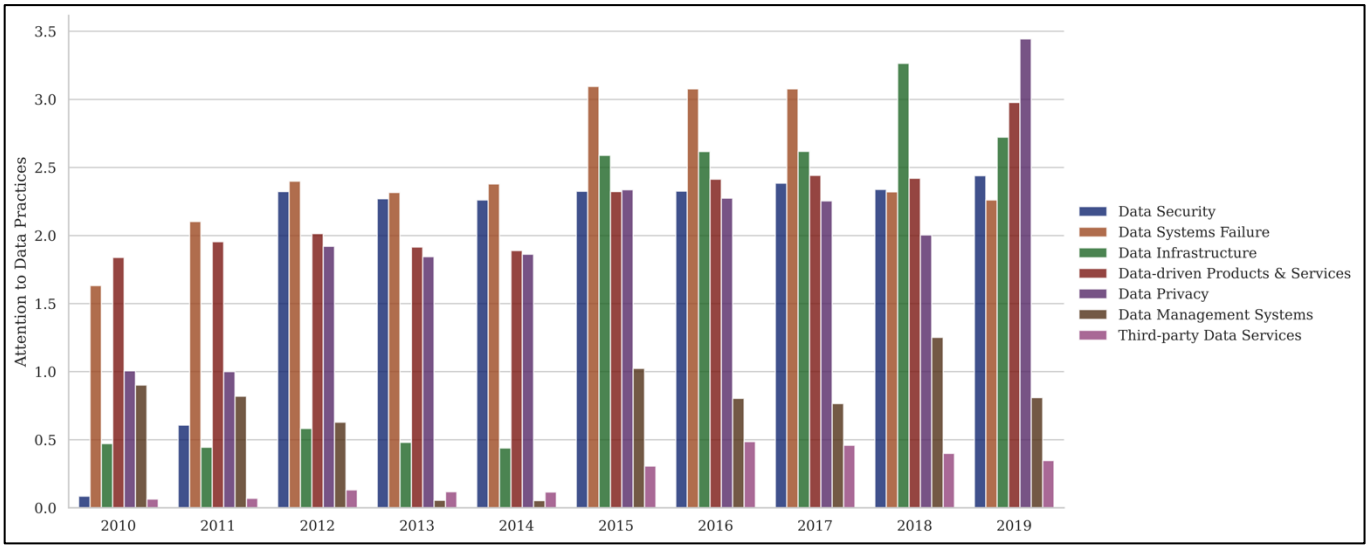


Figure 5: Attention to Data Strategies – Mastercard

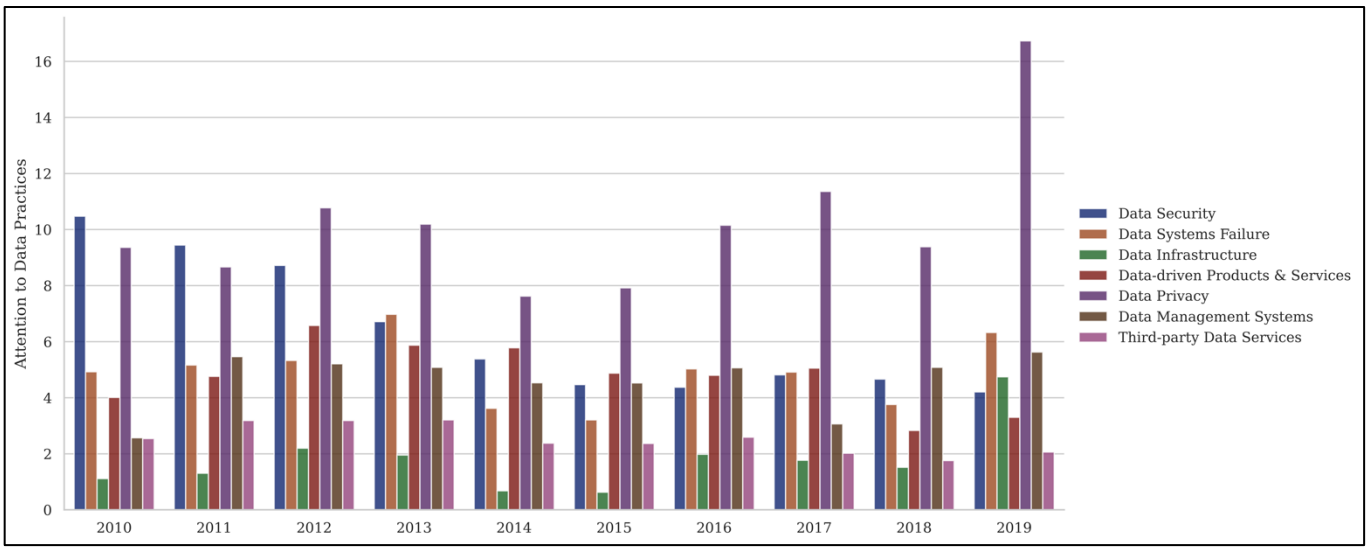
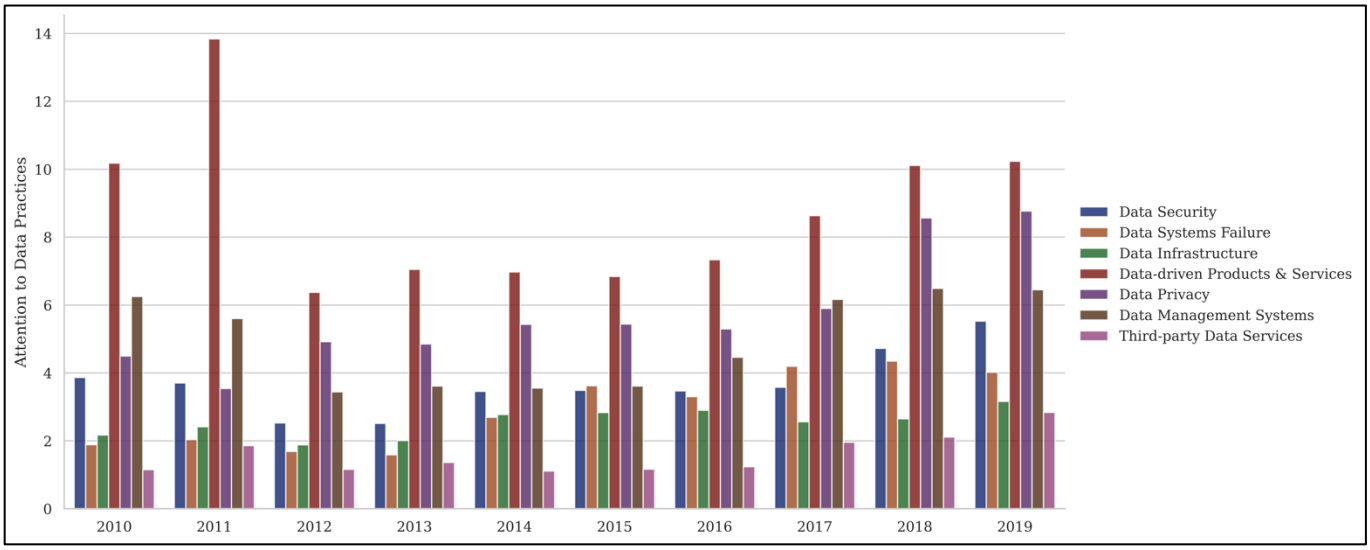


Figure 6: Attention to Data Strategies – United Health



Appendix 3: GDPR Background

The European Union's General Data Protection Regulation (GDPR) regulates the processing of personal data of E.U. residents. The regulation was first introduced to the European Parliament in 2012. In March 2014, the European Parliament, the legislative branch of European Union, approved GDPR. The regulation was adopted in its current form by the European Parliament, European Council, and European Commission in May 2016. Though passed in May 2016, enforcement of the GDPR was delayed until May 25, 3 2018 to allow stakeholders to adjust. The regulation acknowledges the global nature of information flows and applies to both E.U. firms and non-E.U. firms that target E.U. residents. GDPR fines can reach 4% of a firm's annual global revenue.

GDPR is a multifaceted regulation and is built upon seven principles of 1) lawfulness, fairness, and transparency meaning that data processing must be "lawful, fair, and transparent to the data subject", 2) purpose limitation meaning that data processing should be for legitimate purposes specified explicitly to the data subject when data was collected , 3) data minimization meaning that only the data that is absolutely necessary for the specified purpose should be collected, 4) accuracy meaning that the personal data by a data controller should be accurate and up to date, 5) storage limitation meaning that personal data can be stored for as long as the original purpose specified at the time of collection holds, 6) integrity and confidentiality meaning that the collection should be done with integrity, confidentiality, and security, 7) accountability meaning that the data controller is responsible for being able demonstrate compliance with all these principles (GDPR.eu 2020). A key principle of GDPR is data minimization which means that firms are not allowed to collect personal data unless necessary for a specified purpose and that they will not be allowed to process the collected data for any purpose other than the one specified at the time of collection. Firms are required to audit internal data processes, encrypt and anonymize personal data, and notify affected individuals and the regulator in the event of a data breach. Firms are also responsible for respecting the new data rights of E.U. residents under the GDPR, including the right to be forgotten, the right to access personal data, the right to correct data, right to transfer data, and the right to object to data processing (GDPR.eu 2020). In summary, the GDPR requires firms to limit

personal data processing with potential consequences for both its associated operational cost and legal liability.

The GDPR has significant implications for sectors such as technology, healthcare, retail, and finance and their methods of operation. In particular, data defined as personal under the GDPR includes an individual's financial, healthcare, and web browsing data that are used extensively by companies in these sectors. Post-GDPR, companies are restricted from sharing such personal user data with third parties, except under explicitly defined conditions and with data subject's valid consent (GDPR.eu 2020). Valid consent under the GDPR requires that individuals opt in to data processing and that consent notices must list both the purposes of data processing and all third-parties processing the data. Such restrictions limit the application of the collected data for any purpose other than what the data subject explicitly consented (GDPR.eu 2020).

In the three years since the GDPR's implementation, E.U. regulators have released multiple reports critical of industry practices and levied sizable fines, such as EUR 746 million to Amazon in Luxembourg (2021), EUR 285 million to WhatsApp and Facebook in Ireland (2021), and EUR 200 million to Google in France (multiple fines between 2019 and 2021). However, concerns have been raised about unequal enforcement across national jurisdictions, as EU member states are individually responsible for funding and managing GDPR enforcement activities.

Appendix 4: Supplementary Analyses

Table 8: GMM IV Estimation of ML Innovations

	(1) log (ML patents)	(2) log (ML patents)	(3) log (ML patents)
log (size)	-0.375*	-0.606+	-0.736**
	(0.158)	(0.332)	(0.227)
log (cash holdings)	-1.270***	-1.478**	0.0375
	(0.327)	(0.546)	(0.355)
log (R&D)	0.958***	0.971***	0.884**
	(0.150)	(0.200)	(0.330)
Attention to privacy	-0.111*	-0.187**	-0.357*
	(0.0474)	(0.0666)	(0.175)
Asia presence		4.900+	-6.836*
		(2.780)	(2.862)
Asia presence × Attention to privacy			1.937**
			(0.718)
Constant	5.550***	6.521**	0.735
	(1.333)	(2.335)	(1.579)
N	5970	5970	5970
Adjusted R^2			
Firm fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
AR (2)	-1.512	-1.728	-1.935
AR (2) (p)	0.131	0.084	0.053
Sargan test	3.451	1.890	2.920
Sargan test (p)	0.485	0.756	0.232

Standard errors in parentheses. One-step system GMM with Roodman-collapsed GMM-instruments formed of the lagged dependent variable, GDPR, peers' data breaches, and peers' sales in Asia.
+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: GMM IV Estimate of Tobin's Q using Alternative Measure of Technological Capabilities in Digitization

	(1) Tobin's Q	(2) Tobin's Q	(3) Tobin's Q	(4) Tobin's Q	(5) Tobin's Q	(6) Tobin's Q
log(total assets)	11.93*	21.04**	3.910*	6.910***	6.029*	6.054*
	(4.838)	(8.147)	(1.563)	(2.056)	(2.844)	(2.690)
EBITDA Ratio	-0.210*	-0.150	-0.357***	-0.283**	-0.275	-0.414***
	(0.0960)	(0.178)	(0.0705)	(0.105)	(0.193)	(0.119)
CAPEX Ratio	0.130	7.751	-3.557+	-4.551	-3.410	1.453
	(3.818)	(6.889)	(2.098)	(3.222)	(4.991)	(3.521)
log(R&D)	16.19***	11.95***	4.673***	11.10**	12.82+	3.411
	(1.969)	(3.455)	(1.200)	(4.267)	(6.758)	(3.854)
Attention to Privacy	-4.919***	-2.127*	-2.150***	-4.291***	-4.121*	-1.113
	(0.741)	(0.854)	(0.396)	(1.063)	(1.754)	(0.826)
log (Privacy patents)		4.118+	-20.81*			13.32
		(2.411)	(9.888)			(25.02)
log (Privacy patents) × Attention to Privacy			3.614*			-0.0131
			(1.666)			(2.209)
Asia Presence				55.48	-171.6	-69.94
				(44.25)	(116.0)	(52.30)
Asia Presence × Attention to Privacy					29.31*	18.03**
					(12.64)	(5.585)
Constant	-85.22**	-160.0*	-23.79*	-50.17**	-61.91*	-46.00*
	(30.98)	(65.73)	(10.39)	(15.48)	(24.44)	(20.45)
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
AR(2)	0.519	0.969	-0.816	-1.732	-1.525	-0.479
AR(2) (p)	0.604	0.333	0.414	0.083	0.127	0.632
Sargan Test	4.435	6.356	41.960	4.397	1.917	5.457
Sargan Test (p)	0.350	0.607	0.000	0.355	0.993	0.859

One-step system GMM. We formed Roodman-collapsed GMM-instruments from the lagged dependent variable, GDPR, peers' data breaches, and peers' sales in Asia.
Standard errors in parentheses
$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: GMM IV Estimate of Data Breaches

	(1) Firm's Number of Breaches
log(total assets)	-0.101
	(0.0872)
EBITDA Ratio	0.00422
	(0.0299)
CAPEX Ratio	1.371
	(2.027)
log(R&D)	0.198^{+}
	(0.108)
Attention to Privacy	-0.0859*
	(0.0419)
Constant	0.728
	(0.669)
AR(2)	-1.467
AR(2) (p)	0.142
Sargan Test	16.446
Sargan Test (p)	0.058

Standard errors in parentheses. One-step system GMM with Roodman-collapsed GMM-instruments formed of the lagged dependent variable (focal firm's data breaches), peers' data breaches, and GDPR as instruments.

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 7: Estimated Effect Size for Different Counterfactual Treatment Years

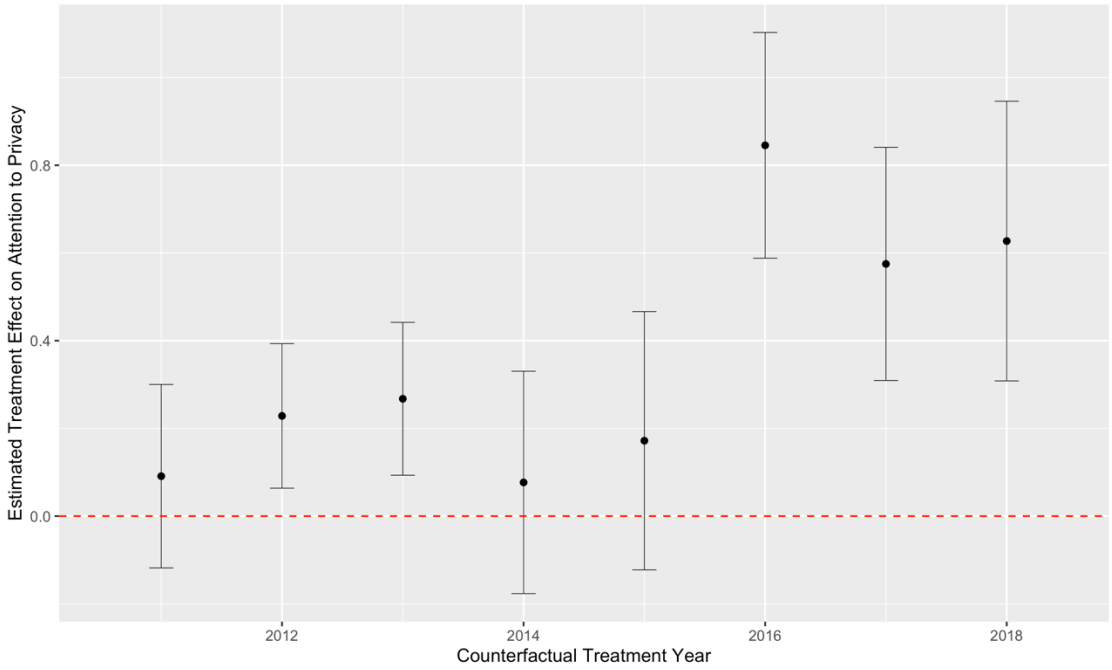


Table 11: Placebo Dependent Variable Tests

	(1) Attention to Data Systems Failure	(2) Attention to Data Infrastructure	(3) Data-driven Products & Services	(4) Data Management Systems	(5) Third-Party Data Services
log(size)	0.529***	0.540***	0.576**	0.536*	0.177**
	(0.115)	(0.164)	(0.197)	(0.239)	(0.0677)
log(R&D)	0.0799	0.544***	0.595***	0.902***	0.195***
	(0.0657)	(0.0936)	(0.113)	(0.137)	(0.0387)
Post GDPR	0.821***	0.117	0.000990	-0.0311	0.309***
	(0.0714)	(0.102)	(0.123)	(0.149)	(0.0421)
EU presence	0.118	-0.0557	0.227	0.0605	0.0400
	(0.0835)	(0.119)	(0.143)	(0.174)	(0.0492)
EU Presence × Post GDPR	0.0252	0.154	0.0177	0.143	0.101+
	(0.0886)	(0.126)	(0.152)	(0.184)	(0.0522)
Peers' data breaches	-0.00210	0.00193	0.00153	-0.00307	-0.000457
	(0.00272)	(0.00388)	(0.00467)	(0.00567)	(0.00160)
Constant	0.837***	0.743***	1.309***	2.223***	0.671***
	(0.153)	(0.218)	(0.262)	(0.318)	(0.0900)
Firm fixed effects	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	3924	3924	3924	3924	3924
Adjusted R^2	-0.113	-0.243	-0.250	-0.245	-0.171

Standard errors in parentheses

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$