Data Risk, Firm Growth and Innovation

Roxana Mihet[*], Kumar Rishabh[†], and Orlando Gomes[‡§]

March 30, 2024

Abstract

In today's modern economy, data stands as a critical asset for firms, yet it is fraught with risks including loss and destruction. In this paper, we examine how data risk impacts firm growth, financial outcomes, and innovation activities. Examining the universe of U.S. publicly listed firms from 2000 to 2022, we find that higher data risk unequivocally reduces knowledge stocks, decreases productivity, and slows growth for the average firm in the U.S. economy. Notwithstanding, there exists a select group of AI-intensive firms, highly exposed to data risk, which develop data protection strategies that enhance innovation and productivity in other domains. This positive spillover leads to higher growth and profitability for these firms. The mechanism is that the same data engineers who develop data protection at these firms are also among the same inventors doing product innovation for these firms. In a second stage, we develop a structural heterogeneous-firm growth model of the data economy to rationalize the empirical findings and provide some counterfactuals.

Keywords: Data economy, data risk, growth, artificial intelligence, innovation.

JEL-Codes: D8, O3, O4, G3, L1, L2, M1.

[*]SFI at HEC Lausanne and CEPR. Contact: roxana.mihet@unil.ch. Corresponding author.

[†]University of Lausanne and University of Basel. Contact: kumar.rishabh@unil.ch.

[‡]Lisbon Accounting and Business School, ISCAL. Contact: omgomes@iscal.ipl.pt.

[§]First version: January 31, 2023. This version: March 30, 2024. We are indebted to Chris Florackis, Christodoulos Louca, Roni Michaely, and Michael Weber for sharing their data on cyber-risk with us. We also thank Elliott Bertrand from Effixis for sharing data intelligence and Francesco Celentano for sharing public cyber-attacks data from Audit Analytics. We also thank our discussants, Fabrice Collard, Luca Sandrini, and Baozhong Yang for their helpful feedback, as well as participants at the EEA 2023, SFI Annual Days 2023, the Economics of ICT 2023, 4th Annual Boca-ECGI, ERMAS 2023, and Toulouse Digital Economics 2024 conferences for useful feedback. We are also grateful for insightful suggestions from Andreas Fuster, Leonardo Gambacorta, Tarun Ramadorai, Norman Schürhoff and Laura Veldkamp. Roxana Mihet acknowledges generous funding for this project from The Sandoz Family Foundation - Monique de Meuron Programme, as well as from UNIL's Dean's Office. All authors declare no conflicts of interest related to this project. All errors are our own.

"In the digital age, your data is your most valuable asset, and protecting it is paramount for any firm because once it's breached, the consequences can be irreversible." Tim Cook (CEO of Apple)

1 Motivation

The modern economy is a data economy. Data not only underpins strategic decision-making, but also acts as a crucial input in AI-driven decision processes that propel firms toward innovation, growth, and competitive advantage (Brynjolfsson et al. (2017), Agrawal et al. (2022) and Babina et al. (2024)). However, this invaluable asset is not without its vulnerabilities; it is beset with risks, including, but not limited to, data breaches and loss. These risks pose substantial threats to operational continuity, financial stability, growth and the innovative capacities of firms. Understanding the multifaceted impacts of data risk is crucial, not only for assessing its direct and indirect financial repercussions, but also for evaluating their effects on innovation, particularly within firms leveraging AI-driven decision processes. The role of data as the backbone of these processes amplifies its significance and highlights the critical need for maintaining data integrity. Additionally, the evolving regulatory environment around data privacy, exemplified by the GDPR and CCPA, highlights the critical need for firms to navigate data risks adeptly (Peukert et al. (2022) and Aridor et al. (2023)). Effective data risk management is essential not just to avoid financial damages, as well as operational and reputation costs, but also to leverage AI investments for innovation and competitive advantage. This emphasizes the intertwined, complex nature of data security, regulatory compliance, and technological advancement.

In this paper, we examine how data risk impacts firm growth, financial decisions, and innovation activities. We empirically examine the impact of data risk on the universe of publicly listed firms in the last two decades, and then build a structural model to rationalize the empirical findings. Our empirical strategy consists of conducting a Poisson regression analysis and a staggered difference-in-difference design to study the relationship between data risk and firm outcomes. To measure firm innovation, we use citation-weighted patents from the USPTO and Kogan et al. (2017). To measure individual firm data risk over time we use the text-based NLP method of Florackis et al. (2023), essentially comparing the risk description in a firm's 10K to the risk description of a set of public firms breached in the following year. We then use this measure to investigate whether companies that are highly exposed to data risk experience significant changes to their financial and innovation outcomes.

We address endogeneity concerns[1] by employing an instrumental variables ap-

[1]The intertwined relationship between data risk and innovation, which mutually influence each

proach that provides a clean source of variation in data risk. Our instrument is the staggered adoption of Data Breach Notification Laws in the United States. Data breach notification laws are regulations that require organizations to inform individuals, regulatory authorities, and sometimes other stakeholders when a security breach has occurred, leading to the unauthorized access, disclosure, or loss of personal data. These laws are designed to protect consumer privacy and enhance corporate accountability by ensuring that affected parties are aware of breaches that may impact their personal information. These laws have been shown to increase firm risk related to data breaches (Boasiako and Keefe (2021); Liu and Ni (2023); and Huang and Wang (2021)). Our strategy is to compare the financial decisions and innovation activities of firms located in early-treated states to those of firms located in late-treated states, taking into account the "forbidden comparison" mis-specification in two-way fixed effects staggered differences-in-differences models (Goodman-Bacon (2021) and Borusyak et al. (2022)).

We narrow down our analysis to consider the impact of data risk on AI-intensive firms because data is a crucial input for these firms. We propose a new measure of firm AI-intensity based on firm business description similarity as computed by Hoberg and Phillips (2016) to a set of firms that file AI patents, as defined by Giczy et al. (2022). The advantage of this measure is that it is constructed at firm-year level with only publicly available data. This implies that it can be constructed over long time-periods and for a variety of public firms belonging to any industry. Moreover, by focusing on the close firms that develop AI patents, we also include firms that do not necessarily file AI patents, but heavily use AI in their day-to-day business operations. Our measure is complementary to Babina et al. (2024), who identify the hiring and stock of AI-skilled labor at the firm-year level using worker resumes and job posting data. Our measure is complimentary in that it identifies the use of AI technologies in day-to-day firm operations, not exclusively a firm's AI-skilled labor share.

Both our direct Poisson estimation and our staggered difference-in-difference method suggest that AI-intense firms experience an increase in profitability and patenting activity in response to an increase in data risk, controlling for a multitude of firm-level characteristics. We also find that non-AI intensive firm profitability and innovation fall, while leverage increases, as expected for the average firm in the economy faced with higher data risk. The robustness check analysis indicates that it is not firm size driving the results, but AI-intensity. Within superstar firms, defined by Autor et al. (2020), it is the AI-intensive large firms that drive the positive results, with no effects for non-AI-intensive superstar firms. Lastly, a sub-sample analysis suggests that the positive

other, complicates the direct analysis, making it challenging to establish a clear cause-and-effect relationship. For example, firms that are more innovative might be more susceptible to data risk to begin with.

spillovers are concentrated in AI-intensive firms in the financial sector and tech firms in the retail sector as defined through the NAICS industry code. We then investigate the mechanism through which AI-intensive firms benefit from increased data risk.

The mechanism through which this positive spillover occurs is that data risk ex-ante prompts AI-intensive firms to pursue digital innovation that enhances productivity in other domains. Data risk forces these companies to improve their data protection measures and systems, which can lead to the development of new technology and products. A concrete example of this positive spillover when facing higher data risk is that, in the pursuit of finding ways to securely store and transmit financial information over internet networks, Amazon used their own-built solution (Amazon's 7th and 9th most cited patents) to offer the new "1-Click ordering" feature (Amazon's most cited patent). Thus, the need to protect firms against data risk has created a demand for skilled labor, and more secure software, hardware, and services, which has lead to technological advancements and to a boost in long-term growth.

Investigating the mechanism further, we find that in AI-intensive firms, the IT department in charge of data protection is highly interconnected with the R&D department in charge of product development. We find that when data risk increases, the share of self data protection patent citations increases dramatically for AI-intensive firms, but does not respond in non-AI-intensive firms. Moreover, we find that the average share of inventors of data security patents, who also work on non-data security patents, is 7-8 times larger in AI-firms relative to their non-AI-intensive peers. We do not observe that AI-intensive firms increase the share of common inventors after data risk increases. Rather, they already have the capacity to employ the same inventors on both data security and non-security patents.

We also perform a sub-sample analysis to understand which industries strongly respond to data risk. Using the NAICS classification, we observe that among AI-intensive firms, which span many different industries, financial firms and retail tech firms respond most positively to an increase in data risk, while firms in the health and accommodation industries are most adversely affected. The observed variance in response to increased data risk among industries classified by the NAICS, notably between the financial, insurance, retail tech sectors, and the health and accommodation sectors, can primarily be attributed to differences in data dependency, regulatory pressures, and inherent adaptability. Financial and tech firms, being inherently data-driven and technologically adept, view data risks as opportunities for innovation and competitive strengthening, underpinned by substantial investments in data security. Conversely, health and accommodation sectors, heavily regulated and less technologically focused, face significant challenges, as heightened data risks translate into operational vulnerabilities and increased compliance costs, without commensurate benefits.

In the second part of the study, we develop a theoretical framework that rationalizes the main mechanisms driving the interaction between data risk and digital innovation. We build a heterogeneous-firm growth model of the data economy, in which data is information that helps firms optimize their business processes and is subject to data risk, meaning that it can be damaged and destroyed. Firms are heterogeneous in their AI-intensity levels and are allowed to protect themselves against data risk. AI-intensity is modeled as firms being able to develop in-house security solutions that are specifically tailored to the needs of firms and investing in data protection has a side-effect of increasing the maximum quality frontier of the goods products. Non-AI intensive firms in the model purchase non-rival data security from AI-intensive firms. Both types of firms solve profit maximization problems and they both benefit from fighting data risk: for the latter, the single benefit of purchasing protection consists in preserving data; for the former, in-house innovation signifies not only the preservation of data, but also an innovation spillover effect that increases the potential quality of the produced goods. We use the simple model to provide some counterfactual analyses.

Our research is nestled in the burgeoning dialogue on data as a critical asset in the empirical and theoretical landscape, illustrating its dual role as a catalyst for strategic decision-making and AI-driven innovation (Babina et al. (2024)) and as a source of risk (Mihet and Philippon (2019)). While the potential of data to propel firms toward growth and competitive advantage is well-acknowledged, the associated risks—ranging from data loss to breaches—pose significant challenges. Our work contributes to the understanding of these dual facets by examining the impact of data risk on firms' innovation, growth, and profitability, offering a nuanced view that goes beyond the conventional focus on firm valuation and equity returns impacted by data risk (Jamilov et al. (2021), Akey et al. (2018), Florackis et al. (2023)).

Furthermore, we explore the dynamic interplay between data risk and data protection within the financial and tech sectors, underscoring its systemic significance (Duffie and Younger (2019), Aldasoro et al. (2022)). By employing novel methodologies to analyze corporate disclosures (Florackis et al. (2023)), our study reveals how AI-intensive firms adeptly navigate the complexities of data risk, not merely mitigating adverse impacts, but also leveraging these challenges to spur innovation and secure competitive advantage. This proactive approach to data risk management showcases a significant amplification of innovation activities, evidencing the critical role of data security in fostering cross-disciplinary innovation and enhancing firm resilience. Our findings not only augment the empirical literature on data risk but also contribute to the theoretical discourse on the role of data in economic growth (Farboodi et al. (2019), Jones and Tonetti (2020), and Eeckhout and Veldkamp (2022)), highlighting the indispensable need for robust data protection strategies in the digital era.

The remainder of the paper proceeds as follows. Section 2 describes the data and hypotheses and tests our main predictions. Section 3 addresses endogeneity and explains our staggered differences-in-differences strategy and results. Section 4 builds a model of the data economy with data risk and protection and provides some comparative statics exercises. Section 5 concludes.

2 Empirical Analysis

The empirical analysis begins by exploring how data risk influences key aspects of firms' operations, including profitability, growth, and innovation activities. We delve into the question of whether heightened data risk leads to a deceleration in firm growth, a reduction in profitability, and a decline in innovation efforts as firms potentially need to allocate substantial resources towards bolstering their data protection measures. This investigation aims to uncover the extent to which the challenges posed by increased data risk divert critical resources and attention from growth and innovation-driven initiatives to data security defenses, thereby possibly impacting the overall performance and strategic direction of firms. Through this analysis, we seek to understand the broader implications of data risk on the competitive landscape and the capacity of firms to sustain and enhance their market position in the face of evolving data risk.

2.1 Data

Data risk. We expand the method from Florackis et al. (2023) to create a firm-year level of data breach risk for the period 2000 to 2022. Florackis et al. (2023) builds data security risk scores based on a textual analysis of the annual 10-K filings of these companies. For any given year, a firm's risk measure is derived from the similarity between the language used to detail *data risk-factors* in its current-year 10-K filings and the *previous-year* 10-K filings of a chosen 'training' set of firms. The firms in this training set are those that endured actual data breaches in the same year. We source the list of breached firms from Audit Analytics. The assumption is that firms that have fallen prey to actual data breaches likely had existing vulnerabilities, which would have been reflected in their risk disclosures in the previous year. As such, if a firm's language in its risk-factor disclosures strongly resembles the previous-year risk disclosures of firms that were indeed attacked, it is inferred it features high data security risks. The similarity score, which also serves as the data risk score, ranges from zero to one, with a higher score indicating a greater data breach risk.[2] We compute these

[2]We believe these are good measures of firm data risk because US firms are required by law to report data breaches in all 50 states, the District of Columbia, Guam, Puerto Rico and the Virgin Islands,

firm-year level scores for the period 2000-2022. The risk index is robust to using a smaller data risk dictionary or no dictionary at all. More details on the exact method of computation can be found in Appendix *A*.

Innovation. We capture innovation in various complementary ways. The first measure we use is the knowledge capital accumulation calculated by Ewens et al. (2020). Knowledge capital is the stock of research and development (R&D) expenditure net of the knowledge capital depreciation. Knowledge asset can also be thought of as an input to innovation, rather than output, as it represents expenditure on producing innovation. Our next set of measures explicitly capture innovation output.

Firms' patent activity represent their innovation output. Following the literature on innovation, we count patents filed by the firms by taking into account their scientific and economic value (Kogan et al., 2017; Aghion et al., 2013; Howell, 2017). In our first patent measure, we count number of patents filed by weighing it with the number of forward citations they receive. The idea is that the more scientifically important a patent is, the more citations it receives (Hall et al., 2005; Kogan et al., 2017). Following the best practice in the literature, we adjust the count for the truncation bias. As the citations occur over time, a simple counting of cites underestimates the importance of the patents that were issued towards the end of our sample period (Lerner and Seru, 2022; Dass et al., 2017; Hall et al., 2001). We correct for that using the well-established methodology proposed by Hall et al. (2001).

We also calculate value-weighted count of number of patents filed. We do so by weighing each patent by the economic value it creates. The economic value of a patent is the dollar amount of wealth generated for the patenting firm's shareholders, calculated from the stock market response to the news about the patent award. We scale the patent value by the firm's total assets, following Kogan et al. (2017).

In an additional analysis, we examine whether firms exposed to more risks expand their areas of innovation. To do that, we extract the Cooperative Patent Classification (CPC) code for each filed patent. We then count the number of unique 'fields' in which a firm files patents in a year. We define number of fields at different level of coarseness. A CPC code consists of five hierarchical parts: section, class, sub-class, group, and sub-group. Section is the highest level in hierarchy, and the most aggregative level, followed by class, subclass, and so on. For our purpose, we define patent fields at three alternative levels: section, class, and sub-class. We do not differentiate patents along the group or subgroup levels because we want to make sure that we are counting patent

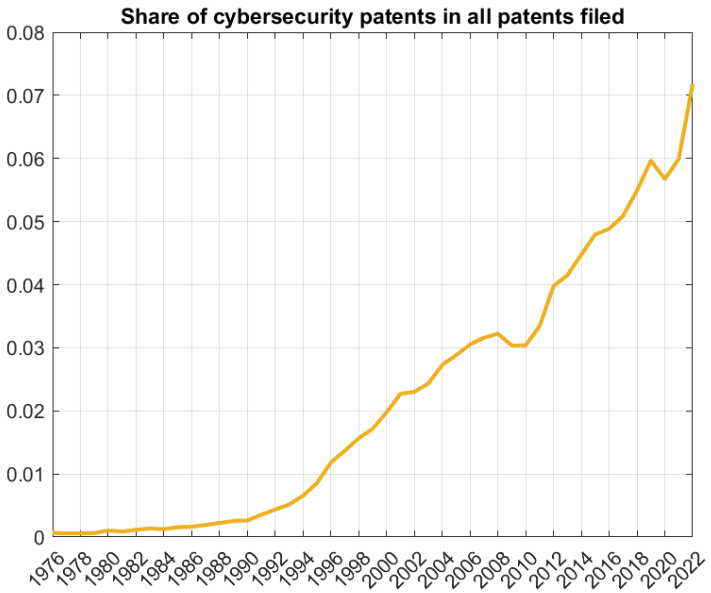
and therefore they are highly likely to be forthcoming about their data risk and risk mitigation in their 10-Ks (Murciano-Goroff (n.d.)). Moreover, we are confident in the validity of this measure because, according to the undertaken calculations, it correlates highly (86%) with the data risk measure based on conference calls from Jamilov et al. (2021) and 95% with our own measure using a smaller set of data-risk related words.

fields that are somewhat distinct from each other.

All our patent data is from the publicly available database maintained by the authors of Kogan et al. (2017).

Data security innovation. We measure cyber security innovation using the citation-weighted and value-weighted count of cyber security patents a firm files within a year. A patent is classified as a cyber security patent if the USPTO assigns it CPC codes associated with cyber security. For instance, CPC code G06F21/ is titled "Security arrangements for protecting computers, components thereof, programs or data against unauthorised activity". Our cyber security patent measure indicates a consistent growth in cyber security innovation over time, currently accounting for approximately seven percent of all patent filings (as depicted in Figure 1).

Figure 1: Data security innovation



AI-intensity score. To identify AI-intensive firms, we create a measure grounded on two fundamental premises. First, we propose that firms involved in the creation of AI technology are, by nature, AI-intensive. Second, we posit that any firm, including those not directly engaged in AI development, can be considered AI-intensive if the language used to describe its business mirrors that of AI-creating firms. In accordance with these premises, our measure is crafted in two steps. For the initial step, we utilize a newly published dataset by the USPTO, a product of their internal research, which classifies AI patents within the entire spectrum of patents filed at the USPTO (Giczy et al. (2022)). This helps us identify those US public firms that have submitted AI patent applications. In the second step, we employ a dataset curated by Hoberg and Phillips (2016), which quantifies the textual similarity in the business description between any

two firms. The underlying principle here is that AI-intensive firms are likely to portray their businesses in a similar light. Therefore, a firm not holding an AI patent is also considered AI-intensive if its business description more closely aligns with those firms possessing AI patents.

Financial variables. We obtain firm level financial information from the merged CRSP-Compustat database. We calculate various financial variables and ratios to use them as control variables in our baseline regressions. Specifically, we use the following variables as controls: log of total assets, Tobin's Q, asset tangibility, book-to-market ratio, cash-to-asset ratio, leverage, and return on assets. We winsorize all the variables at 0.5% on both sides of the distribution.

Table 1 presents summary statistics on our data risk and innovation measures. We see that more than a quarter of the firms do not face data risk. Further, as is well-known innovation activity is quite skewed. For instance, more than 50 percent of firm-years do not record any positive knowledge capital accumulation or any patent activity.

Table 1: Summary statistics on cyber-risk score and innovation variables

	N	mean	sd	p10	p25	p50	p75	p90	p99
Cyber-risk score	44972	0.2	0.2	0	0	0.3	0.4	0.5	0.6
Log(Knowledge stock)	41479	1.6	2.2	0	0	0	3.4	5.0	8.0
Log(R&D expenditure)	44972	1.3	1.9	0	0	0	2.6	4.2	7.2
Patents filed: simple count	44972	9.2	49.9	0	0	0	0	8.0	291.0
Patents filed: citation-weighted count	44972	18.3	100.4	0	0	0	0	15.3	549.4
Patents filed: value-weighted count	44881	0.05	0.20	0	0	0	0	0.11	1.17
Number of patent sections	10616	3.3	2.0	1	2	3	4	6	9
Number of patent classes	10616	7.5	10.3	1	2	4	8	17	57
Number of patent subclasses	10616	14.1	25.0	1	3	6	14	31	145

N refers to the total number of firm-year. p10-p99 refer to the 10th to 99th percentile values. Cyber risk score lies between zero and one, with higher values indicating higher risk. Cyber risk score measure is obtained from Florackis et al. (2023). Knowledge stock is based on the estimates of knowledge stock net of knowledge depreciation from Ewens et al. (2020). Simple patent count refers to number of patents filed by the firm in a year. Citation-weighted patent count weighs each patent with the forward citation the patent receives, adjusting for the filing vintage. Value-weighted patent count is the sum of stock market value generated over all the patents filed by a firm in a year, scaled by total assets. Number of patent sections refers to the number of unique CPC sections associated with all the patent the firm files in a year. Similar explanation applies to patent classes, and subclasses, respectively.

2.2 Empirical strategy

We first conduct regression analysis to uncover the reduced-form correlation between data risk and firms' financial, growth, and innovation strategies. We rely on two aspects of our regression specification. First, we regress firm outcomes on the lagged value of data risk score. Doing so addresses the simultaneity concerns. Second, we include firm fixed effects to absorb time invariant characteristics of firms that might

affect this relationship. Moreover, we include year fixed effects to absorb shocks occurring over time and that are common across firms. Finally, we control for a multitude of financial firm characteristics.

As visible from Table 1, the firm outcomes, particularly the innovation variables, have a right skew and contain a high share of zeros. Therefore, applying ordinary least squares (OLS) estimation would result in inefficient parameter estimates. While there are some possible solutions that transform the firm outcome variables, these methods are known to produce inconsistent and incorrect estimates.[3] Thus, we use the Poisson model to explicitly take into account many zeros and the right skew of the dependent variables, as recommended by (Cohn et al., 2022) and (Correia et al., 2020).[4]. Other studies employing Poisson regression with patent data include (Azoulay et al., 2019; Aghion et al., 2013; Amore et al., 2013; Blundell et al., 1999; Hausman et al., 1984).

To study the relationship between the lagged value of data risk score (data risk_{it-1}) and firm outcomes (y_{it}) we fit the following conditional expectation of the firm outcome:

$$\mathbb{E}[\text{y}_{it}|\text{data risk}_{it-1}, \mathbf{x}_{it-1}, \eta_i, \tau_t] = \exp\left(\beta_c \text{data risk}_{it-1} + \beta \mathbf{x}_{it-1} + \eta_i + \tau_t\right) \tag{1}$$

where y_{it} is the firm's innovation variables such as citation-weighted patent counts, knowledge, R&D, and financial variables such as log assets and return on assets. data risk_{it-1} is the lagged value of the data risk score, $\mathbf{x}_{it-1}$ are *lagged* control variables, including log R& D expenditures, size (log of total assets), Tobin's Q, asset tangibility, book-to-market ratio, cash-to-asset ratio, leverage, and return on assets, when these controls are not the dependent variable itself. We denote η_i as the firm fixed effect, and τ_t as the year fixed effect.

We perform Poisson pseudo-maximum likelihood estimation to estimate the parameters of the model in (1). Finally, we cluster standard errors at the firm level, to take into account the possibility of autocorrelation and hetereskedasticity in the error terms. Clustered standard errors are additionally useful because they are also robust to 'overdispersion' (and 'underdispersion') issues countered in Poisson regression (Cohn et al., 2022; Wooldridge, 1999).

[3]One approach is to use OLS estimation with a log transformation of the patent count variables, although the presence of many zeros means this method might exclude a significant number of observations. Log-linear regressions, while a potential solution, are criticized for potentially yielding inconsistent parameter estimates. An alternative could involve adding 1 to each patent count before transformation or employing the inverse hyperbolic sine transformation. These methods keep zero counts but might still lead to inconsistent estimates and could misrepresent the true relationship's direction (Cohn et al. (2022), Silva and Tenreyro (2006)).

[4]The Poisson model offers consistent estimators for count data, like patents, without assuming higher order error moments. It also supports group fixed effects, crucial for our analysis. Notably, the model is applicable not only to discrete count data but also to continuous non-negative variables, such as knowledge assets (Silva and Tenreyro (2011), Wooldridge (1999)).

2.3 Baseline results

Firm outcomes. Does a rise in data risk affect a firm's profitability? Does it decrease a firm's innovation activity? Data risk can reduce firms' growth and innovation outcomes by diverting its resources towards data risk protection measures.

Table 2 presents the results from our preferred Poisson estimation of regressing firm outcomes on lagged data risk. We find that firms file more overall patents, more non-data security patents, accumulate more knowledge capital and R&D stock in response to a rise in data risk, and increase their size and profitability. In Appendix B, we show the results are robust to using a battery of different controls.

Table 2: Regression of firm innovation, growth, and profitability outcomes

	Citation-weighted Patent Counts		Knowledge and R&D		Financial Vars	
	(1) Overall patent count	(2) Non-CS patent count	(3) Knowledge	(4) R&D	(5) Log assets	(6) ROA
L. Data-risk score	0.243**	0.226*	0.0612	0.122*	0.159**	0.065***
	(0.134)	(0.131)	(0.0563)	(0.0683)	(0.060)	(0.019)
L. Firm Controls	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	12900	14122	15111	21358	20238	20234

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are in parentheses. Standard errors are clustered at the firm level. N refers to the total number of firm-year. Data-risk score, and all control variables are lagged by one year. The data-risk score is computed as explained in A. Knowledge is based on the estimates of knowledge stock net of knowledge depreciation from Ewens et al. (2020). Other control variables are computed using WRDS CRSP-Compustat merged data. R&D expenditures are research and development expenditures typically reported in the income statement, Tobin's Q is defined as Total assets (at) minus common equity (ceq) plus market value of equity (prcc_f × csho), as a ratio of total assets (at). ROA is defined as operating income before depreciation (oibdp) to total assets (at). Tangibility is defined as total property, plant and equipment (ppent) scaled by total assets (at). Leverage is long-term debt (dltt) plus debt in current liabilities (dlc), as a ratio of total assets (at). Book-to-market ratio is book value of common equity (ceq) divided by the market value of common equity (prcc_f × csho). Cash-to-asset is the ratio of cash and short-term investments (che) to total assets (at). For a detailed variable description see table 1.

The increase in innovation and profitability measures are significant. For example, a one standard deviation change in data risk is associated with an increase in the overall patent count of about 7%[$= 0.22(e^{0.243} - 1)$], and to an increase in non data-security related patents of about 5.5% [$= 0.22(e^{0.226} - 1)$]. Do firms file more patents because they accumulate more R&D stock, or do they also respond by increasing their R&D productivity? To test that, we regress patent-count variables on lagged data-risk in columns (1) and (2) including the stock of R&D capital as an explanatory variable. Therefore, the coefficient on data-risk score gives us the estimate of how in response to an increase in data risk, a firm's patent count changes keeping its innovation input (R&D capital) unchanged. Moreover, in column (3) and (4) we examine the response of the intangible knowledge stock as defined by Ewens et al. (2020) and R&D expenditures to an increase in data risk. Although, in the regression of knowledge capital, the

Poisson model does not give a significant coefficient for data risk at the conventional 10% significance level, the value is quite close. Moreover, the results are also confirmed by the regression of R&D stock, which shows a significant rise. The increase is also economically meaningful. For instance, a one standard deviation change in data risk is associated with an increase in R&D of about 3% $[= 0.22(e^{0.124} - 1)]$, keeping everything else the same. Lastly, a one standard deviation change in data risk would lead to an increase in firm size by 3.7%$[= 0.22(e^{0.159} - 1)]$ and to an increase in profitability by 1% $[= 0.22(e^{0.065} - 1)]$.

Firm innovation details. How do firms respond with their patenting output when they face a higher data risk? Which type of patents do they file more? To verify this, we pinpoint product patents filed by the firms in our sample. Product patents symbolize both the genesis of new products and enhancements in the quality of existing ones (Babina et al., 2024). Utilizing the patent claims dataset shared by Ganglmair et al. (2022), which classifies patent claims into product and process claims, we label a patent as a product patent if 50 percent or more of its claims are designated as product claims (Babina et al., 2024), and similarly define process patents.

Table 3: Regression of citation-weighted patent count by patent classification

	Citation weighted count of:			Share of product patents in
	All patents	Product Patents	Process Patents	citation-weighted count
	(1)	(2)	(3)	(4)
L. Data-risk score	0.243*	0.208*	0.0154	0.102**
	(0.134)	(0.113)	(0.133)	(0.0428)
L. Firm controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	13375	11497	10786	8298

$^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. The coefficient estimates are derived from the Poisson pseudo-maximum likelihood estimation. Standard errors are denoted in parentheses and are clustered at the firm level. Here, N represents the total number of firm-year observations. Cyber score, and control variables are lagged by one year. Cyberrisk score is constructed using the methodology Florackis et al. (2023). Citation-weighted patent count weighs each patent with the forward citation the patent receives, adjusting for the filing vintage. Non-cybersecurity patent is any patent not clasified as a cybersecurity patent. A patent is classified as a cybersecurity patent if the USPTO assigns it CPC codes associated with cybersecurity. For instance, CPC code G06F21/ is titled "Security arrangements for protecting computers, components thereof, programs or data against unauthorised activity". Product patents are defined as those having at least 50% of their claims categorized as product claims according to Ganglmair et al. (2022). Similarly, Process patents have more than half of their claims classified as process claims. Share of product patent represents the proportion of citation-weighted count of product patents in the total citation weighted-patent count, which includes both product and process patents. Control variables are lagged and include: Log of total assets, log of R&D expenditure, Tobin's Q, Return On Assets, Tangibility, Leverage, Book-to-market ratio, Cash-to-asset ratio. For variable description see table 1.

We examine whether an escalation in data risk triggers an uptick in product patenting. Our testing methodology is twofold: initially, we regress product patent counts, and subsequently, we regress the ratio of product patent counts to process patent counts. Table 3 unveils the results of these regressions. Column 2 affirms that a one standard

deviation increase in data risk catalyzes around a 5% increase in product innovation in terms of citation-weighted patent counts. We see that the effect is indistinguishable from zero for process patents. Further, we scrutinize whether the surge in product innovation supersedes process innovation by analyzing the proportion of product patent counts in the aggregate. Column 4 reveal that the fraction of product patents indeed ascends as firms confront elevated data risk.

AI-intensive firms. Next, we study how the above dynamics differ between the AI-intensive and non-AI intensive firms. In the modern economy, firms are reliant on data and AI-intensive firms in particular are reliant on big data, which is subject to the risk of being encrypted and possibly destroyed. These AI-intensive firms are also highly innovative and have incentives to protect their data, their most valuable asset.

We construct a dummy variable that takes value 1 if the firm is identified as a AI intensive firm by our earlier described method. We then run the regressions similarly to those in the previous section, however now we interact the lagged data risk score with the dummy on AI-intensity. The results are presented in Table 4.

We find that even though the AI intensive firms account only for a minority of the observations (roughly 38%), our baseline results are driven by them. Indeed, the regressions show that data risk score has even sometimes negative effects on innovation in the non-AI intensive firms, although, the results are never significant.

Table 4: Regression of firm outcomes by AI-intensity

	Citation-weighted Patent Counts				R&D	Financial Vars	
	(1) Overall patent count	(2) Product patent	(3) Process patent	(4) Share of product	(5) R&D	(6) Log assets	(7) ROA
L. Data-risk score×$(AI = 0)$	0.216	0.132	-0.101	0.0745*	0.0783	0.0798	0.0189
	(0.164)	(0.144)	(0.165)	(0.043)	(0.0888)	(0.0509)	(0.0174)
L. Data-risk score×$(AI = 1)$	0.384**	0.347**	0.161	0.159*	0.198*	0.249***	0.0811***
	(0.174)	(0.148)	(0.165)	(0.069)	(0.0816)	(0.070)	(0.027)
L. Firm Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	13375	11497	10786	8298	21358	20238	20234

$^{*}p < 0.10$,$^{**}p < 0.05$,$^{***}p < 0.01$. Standard errors are in parentheses. Standard errors are clustered at the firm level. N refers to the total number of firm-year. Data-risk score, and all control variables are lagged by one year. The data-risk score is computed as explained in A. Knowledge is based on the estimates of knowledge stock net of knowledge depreciation from Ewens et al. (2020). Other control variables are computed using WRDS CRSP-Compustat merged data. Log R&D expenditures are logged research and development expenditures typically reported in the income statement, Tobin's Q is defined as Total assets (at) minus common equity (ceq) plus market value of equity (prcc_f × csho), as a ratio of total assets (at). ROA is defined as operating income before depreciation (oibdp) to total assets (at). Tangibility is defined as total property, plant and equipment (ppent) scaled by total assets (at). Leverage is long-term debt (dltt) plus debt in current liabilities (dlc), as a ratio of total assets (at). Book-to-market ratio is book value of common equity (ceq) divided by the market value of common equity (prcc_f × csho). Cash-to-asset is the ratio of cash and short-term investments (che) to total assets (at).

3 Addressing endogeneity

3.1 Staggered DiD using Data Breach Notification Laws

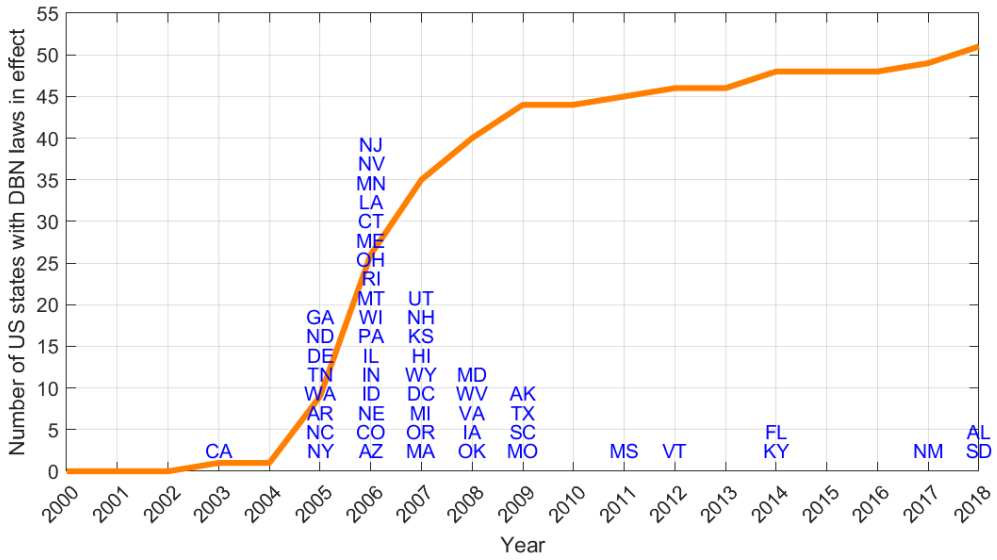
In the previous section we investigated the direct impact of an increase in data risk on firm innovation activities. However, there exists an endogeneity problem in assessing the impact of data risk on firm innovation due to intertwined relationships where data risk and innovation mutually influence each other, making it challenging to establish a clear cause-and-effect relationship. For instance, while data risks might hinder innovation by diverting resources toward data security measures, innovative activities within a firm could also lead to increased data risks due to new technologies or processes being introduced. Moreover, firms that are more innovative might invest more in advanced technologies, making them both more susceptible to data risks and more likely to innovate. This bias can create a spurious relationship between data risk and innovation if not properly addressed. Factors such as reverse causality, omitted variables, simultaneity, and sample selection bias complicate the distinction between the effects of data risk on innovation and vice versa.

In this section, to address this issue, we employ an instrumental variables approach to disentangle and understand the true causal impact of data risk on firm innovation activities. Our instrument is the adoption of Data Breach Notification Laws in the United States, which have been shown to increase firm risk related to data breaches (Boasiako and Keefe (2021); Liu and Ni (2023); Huang and Wang (2021)).

Data Breach Notification Laws (DBNL) in the United States mandate firms to inform individuals affected by a data breach involving their personal information. Typically, these laws require companies that experience a data breach to notify affected individuals within a specified time-frame, often ranging from 30 to 90 days after the breach is discovered. The notification usually includes details about the nature of the breach, the type of information compromised, and steps individuals can take to protect themselves. Additionally, some states require organizations to notify state authorities or consumer reporting agencies depending on the scale and severity of the breach. The laws also have provisions outlining penalties for non-compliance, aiming to hold organizations accountable for safeguarding individuals' personal data. All 50 states have enacted their own versions of DBNL starting in 2003 with California and ending in 2018 with Alabama and South Dakota. By 2008, more than half of the states had adopted a DBN law, as shown in Figure (2).

Our empirical strategy is to explore the staggered implementation of Data Breach Notification Laws (DBN laws) in the United States, which increased firm data risk costs, and compare the innovation activities of firm headquarters located in early-treated

Figure 2: State adoption of DBN laws



Legend: This figure reports the first time in each state and district that a data breach notification law is enacted specifically containing data security breach notification provisions. For example, Nevada introduced a data breach statute in January 2005, but it only required notification provisions for general data provisions in January 2006; thus, in our sample, it appears as a 2006 adoption of DBN law. Only in Nevada also is the ability to launch a private action (2005) different from the date of DBN law adoption. Other states that allow for a private cause of action are: Alabama, Alaska, California, Delaware, D.C., Hawaii, Idaho, Illinois, Louisiana, Maryland, Massachusetts, Minnesota, Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Oregon, Rhode Island, South Carolina, Tennessee, Texas, Washington, and Wisconsin. The source of the data is Perkins Coie LLP (2023).

states to those of firm headquarters located in late-treated states. We must mention that multi-state firms may be affected earlier than the state of their headquarter. If a firm operates in both treated and untreated states, the policy impact might spill over from treated to untreated states via the firm's internal policies, practices, or economic activities, potentially contaminating the control group. Multi-state firms might adjust their operations across states in response to the laws, such as reallocating resources to more favorable regulatory environments, which could affect the estimation of the law's impact. However, these concerns work in our direction, in the sense that the estimation provides a lower-bound for the impact of DBN laws for treated firms relative to a control group which may have already reacted before. While no instrument is perfect, we believe we adequately capture firm responses to an increase in data risk costs because we estimate lower-bounds. In an ideal world, these state laws would have affected only firms in the respective state only, in which case we would have been able to estimate the true average effect, and not the lower bound.

We also make sure we take into account the latest critiques in the literature on staggered difference-in-difference estimation. A very recent literature (Baker et al. (2022);

Goodman-Bacon (2021), among others) has uncovered two vital econometric issues in standard staggered difference-in-difference methods such as linear two-way fixed effects (henceforth, TWFE): (1) there is a possibility of bias due to "forbidden comparisons", and (2) there is a possibility of bias and/or inefficiency due to mis-specification in the presence of right skewed dependent variables. For example, related to the first issue, standard dynamic two-way fixed effects methods suffer from a problem that it aggregates treatment effects over some valid comparisons but also over some "forbidden" comparisons. Specifically, it also compares already treated units (as controls) with the later treated units (as treated). When the treatment effects are heterogeneous over time or across treatment units, it may lead to biased average treatment effect in the treated (ATT) estimates. The second issue of mis-specification in the presence of right skewed dependent variable is also a serious issue. Using a $log(1+y)$ transformation of the dependent variable, a log-linear, or an inverse hyperbolic sine (IHS) regression produces inconsistent and biased estimates. Another method to reduce skewness, the negative binomial regression, does not work with fixed effects.

This leaves us with three models that admit fixed effects and produce unbiased estimates: linear, Poisson, and rate regression. The literature has shown that the Poisson regression is the best because it is the most efficient, having the lowest variance among these three unbiased strategies. Linear regressions can be admitted, however, in spite of problems of high variance, because there are no issues of bias and inconsistency (Cohn et al. (2022)). This will make it harder to get significant results, but at least the estimates will be unbiased and consistent with correct sign. Positive significant results will suggest that *despite* the method producing high variance estimates, there is evidence of an effect of data breach notification laws on firm innovation activities.

In our analysis, we use the Borusyak et al. (2022) linear method (henceforth, BJS) that addresses the first challenge of "forbidden comparisons" and is unbiased and consistent (Cohn et al. (2022)), despite being inefficient. Other popular methods that account for "forbidden comparisons" are Callaway and Sant'Anna (2021), Sun and Abraham (2021), and de Chaisemartin et al. (2020), among others. The BJS estimator is the most efficient under the assumption of parallel trends because it uses all of pre-treatment data in estimation and it is robust to cases when treatment effects vary arbitrarily. The first estimation that we run is a linear difference-in-differences regression accounting for "forbidden comparisons" using the BJS 3-step imputation representation for the efficient estimator, explained henceforth:

1. Within the untreated observations only, estimate the λ_i and δ_t (by $\hat{\lambda}_i^{\star}$, $\hat{\delta}_t^{\star}$) by OLS in

$$Y_{it} = \lambda_i + \delta_t + \epsilon_{it}, \tag{2}$$

where λ_i is unit (i.e. firm) fixed effect, δ_t is year fixed effect;

2. For each treated observation with $w_{it} \neq 0$, set $\hat{Y}_{it} = \hat{\lambda}_i^{\star} + \hat{\delta}_t^{\star}$ and $\hat{\tau}_{it}^{\star} = Y_{it} - \hat{Y}_{it}(0)$ to obtain the estimate of τ_{it};

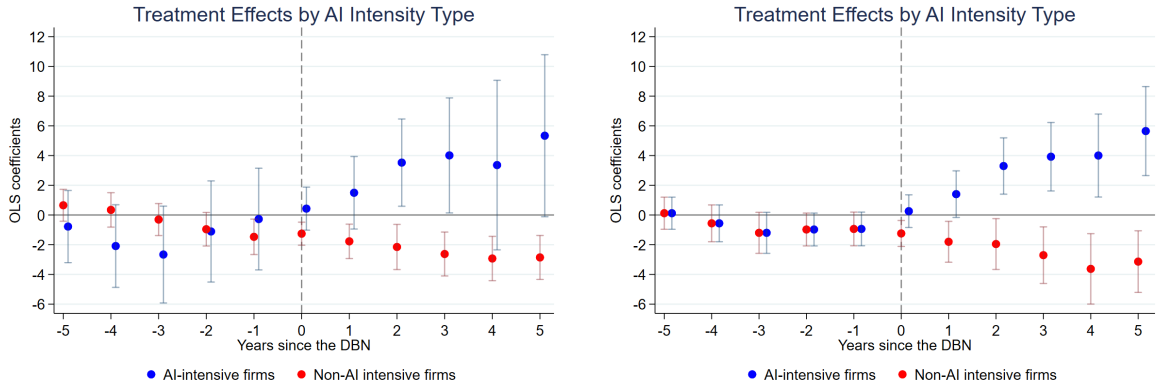
3. Estimate the target τ_w by a weighted sum $\hat{\tau}_w^{\star} = \sum_{it} w_{it} \hat{\tau}_{it}^{\star}$;

The above model allows us to estimate unbiased and consistent dynamic treatment effects using panel data on firms i over years t, where Y_{it} is the time t firm-level measure of innovation, $Y_{it}(0)$ is the period-t stochastic potential outcome of unit i if it is never treated, $\Omega_1 = \{it \in \Omega | treated = 1\}$ is the set of treated observations (i.e., firms are headquartered in a state that has adopted a DBN law), $\Omega_0 = \{it \in \Omega | treated = 0\}$ is the set of untreated (i.e., never-treated and not-yet-treated) observations, $\tau_{it} = \mathbb{E}[Y_{it} - Y_{it}(0)]$ represents the causal effects on the treated observations $it \in \Omega_1$, w_{it} are BJS-derived pre-specified non-stochastic weights that depend on treatment assignment and timing, but not on realized outcomes.

3.2 AI-intensive firms benefit from higher data risk

We explore the impact of Data Breach Notification Laws (DBN laws) on firm innovation activities, as proxied by their patenting activity, as well as on firm growth, cost structure, and profitability.

Figure 3: Citation-weighted patent count by AI intensity



This figure plots BJS-weighted dynamic heterogeneous treatment effects of citation-weighted patent counts by firm AI-intensity pre- and post- treatment. The '0' event is the staggered adoption of DBN laws across the United States. AI intensive firms are identified using a combination of the USPTO dataset on AI patents (Giczy et al. (2022)) and the KPSS patent dataset linked to firms (Kogan et al. (2017)). Moreover, firms that are close to AI patenting firms in the sense of Hoberg and Phillips (2016) and mirror their AI innovations are also considered AI-intensive. This measure has the advantage to be constructed from entirely publicly available data and it is different from IT expenditures.

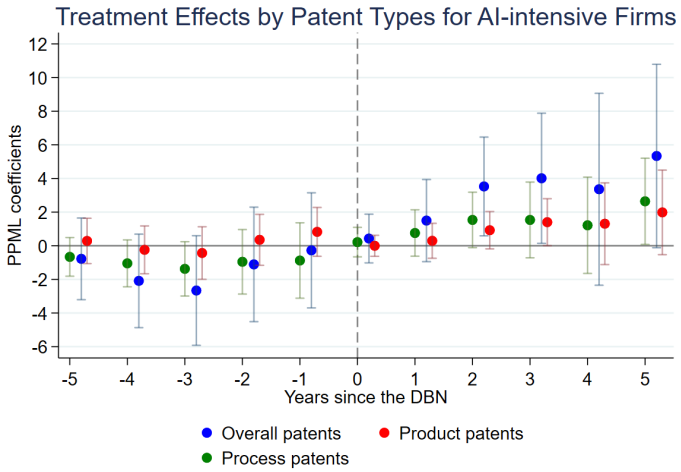
AI-intensive firms patent more. Figure (3) presents the BJS-weighted dynamic heterogeneous treatment effects of citation-weighted patent counts by firm AI-

intensity (AI). The left-hand panel allows heterogeneous pre-trends, while the right-hand panel assumes common pre-trends for both groups, but estimates ATT separately post-treatment.

As shown in Figure (3), AI-intensive firms exhibit higher overall innovation after the adoption of DBN laws. On the other hand, non-AI intensive firms exhibit lower overall innovation after the adoption of DBN laws, suggesting the adoption of these laws is a significant negative shock for firms that imposes high costs which overall discourage innovation. It is to be noted that both panels provide evidence on the observed counterparts of the parallel trends assumption and show that we do not have an unnatural experiment.

Patent class: process vs. product. We also examine differences in patent type (product vs. process patents) after the adoption of DBN laws. We refocus on the product patents filed by the companies in our study. Product patents signify both the introduction of new products and enhancements in the quality of existing ones (Babina et al. (2024)). To identify these patents, we utilize the patent claims dataset provided by Ganglmair et al. (2022), which categorizes patent claims into product and process claims. We classify a patent as a product patent if 50 percent or more of its claims are specified as product claims, following the method described in Babina et al. (2024). Similarly, we establish the definition of process patents in a similar manner.

Figure 4: Citation-weighted patent count: by patent type, for AI-intensive firms



This figure plots the effects of citation-weighted cybersecurity patent counts by firm AI-intensity pre- and post- treatment. The '0' event is the staggered adoption of DBN laws across the United States. Estimates for AI intensive firms are in blue, while estimates for non-AI intensive firms are in red. AI intensive firms are identified as described previously.

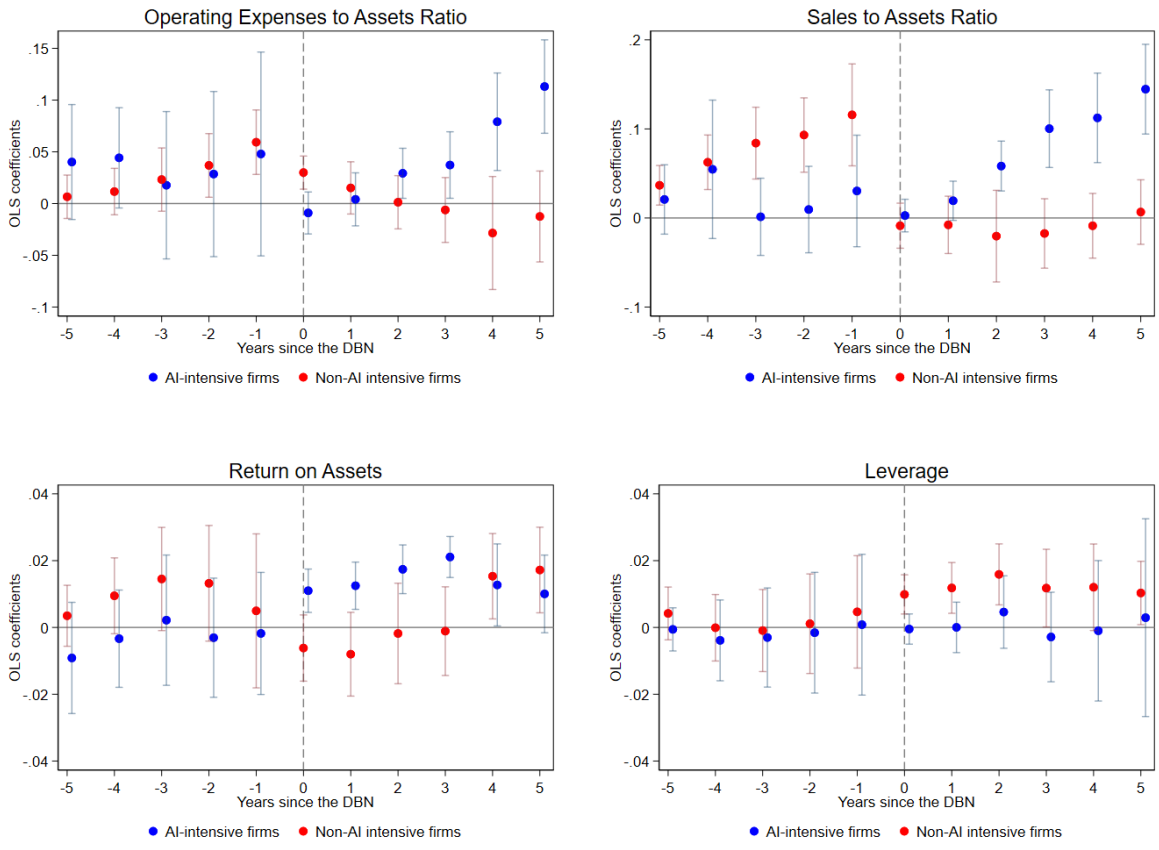
Figure (4) presents the BJS-weighted dynamic heterogeneous treatment effects of

citation-weighted patent counts by patent type (i.e., overall, product, and process) for AI-intensive firms. The left-hand panel allows heterogeneous pre-trends, while the right-hand panel assumes common pre-trends for both groups, but estimates ATT separately post-treatment.

As shown in Figure (4), AI-intensive firms exhibit a slight increase in process and product patenting, although the results are not different or significant except for long-term horizons. As mentioned previously, linear staggered difference-in-difference methods that account for "forbidden comparisons" produce consistent and unbiased estimates, but they may produce insignificant estimates due to high variances, when the dependent variable is right skewed, which is typical of the patent counts.

AI-intensive firms become more profitable and grow more. We repeat the analysis, this time looking at the response of firm costs (operating expenses), size (sales), profitability (return on assets), and leverage to the staggered implementation of Data Breach Notification laws.

Figure 5: Financial variables by AI intensity



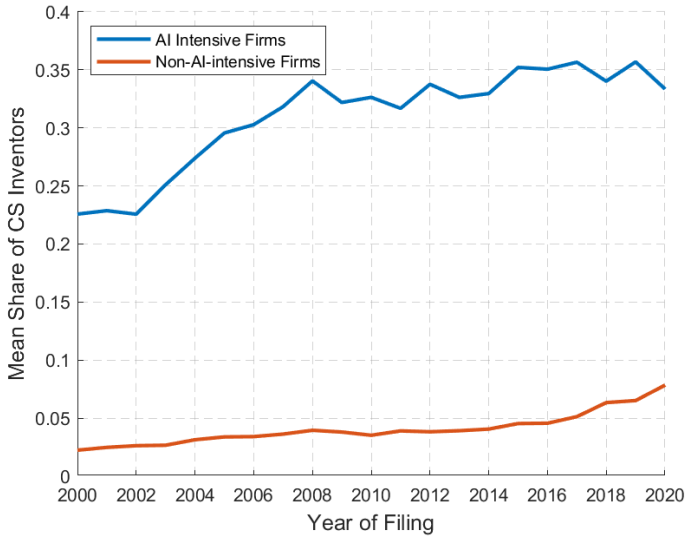
This figure plots BJS-weighted dynamic heterogeneous treatment effects of select financial variables. The '0' event is the staggered adoption of DBN laws across the United States. Return on assets is defined as ratio of Operating income before depreciation to total assets. Leverage is defined as long-term debt + debt in current liab to total assets ratio.

Neither the cost nor the size responses are to be trusted, because of the strong presence of pre-trends. We include them because it is important to be transparent about which results hold in the causal estimation and which do not. On the other hand, we can confidently observe that AI-intensive firms become more profitable after the DBN laws. Moreover, they do not increase their leverage, while non-AI-intensive firms significantly increase their leverage to survive after the adoption of DBN laws.

3.3 Mechanism: inventor network and in-house advantage

Common inventors. In Figure 6, we observe that the share of data security inventors on non-data security patent teams is consistently 7-8 times larger in AI-intensive firms relative to non-AI-intensive firms. This indicates a significant cross-pollination of expertise and innovation between data security and other technological domains within firms that heavily utilize AI. This phenomenon suggests that AI-intensive firms not only prioritize data security to protect their data-rich environments but also leverage the specialized knowledge of data security professionals to enhance innovation across different areas of their business. It could imply that these firms recognize the strategic value of integrating data security insights into broader product development and innovation processes, leading to more robust and secure technological solutions.

Figure 6: Common inventor network by AI intensity



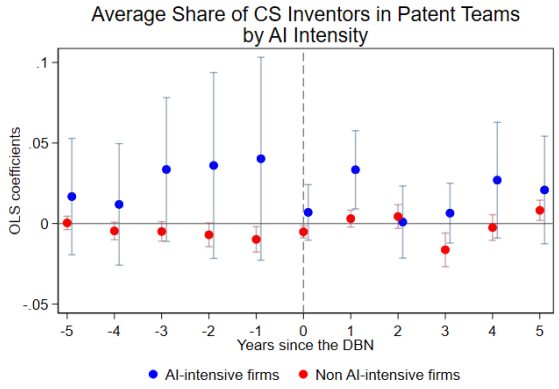
This figure plots the proportion of cybersecurity (CS) inventors within non-CS patent teams, calculated at the patent level and averaged annually across AI-intensive and Non-AI-Intensive firms. Cybersecurity inventors are defined as those who have contributed to the development of at least one cybersecurity-related patent for the same assignee firm as the non-cybersecurity patent in question. Cybersecurity patents are identified through CPC classification codes as assigned by the USPTO as described in the main text.

This interdisciplinary collaboration likely fosters a culture of innovation that is attuned to the complexities of the digital age, where security and functionality are increasingly intertwined. It also points to the potential for AI-intensive firms to drive industry standards and practices in data security, setting benchmarks that could influence the wider market, including non-AI intensive firms.

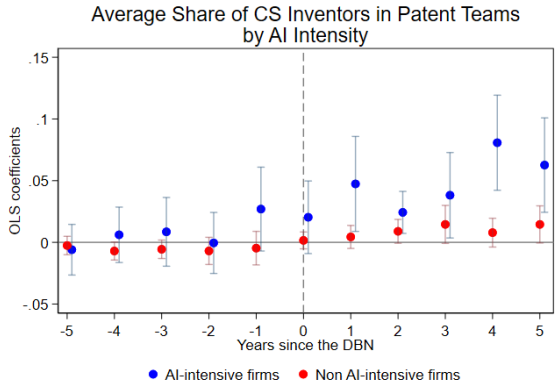
Existing capability or developing new capabilities? We also investigate whether AI-intensive firms already have the capability to innovate both in the data security space and the product/process space, or whether they increase their share of common inventors significantly after the DBN laws. The evidence suggests that AI-intensive firms slightly and temporarily increase the share of common inventors on both data security (CS) and non-data-security patents. The effect does not seem to be persistent for the full sample, although it is more pronounced for superstar firms.

Figure 7: Share of cybersecurity innovators in patent teams

(a) Full Sample



(b) Top 500 firms (2000 vintage)

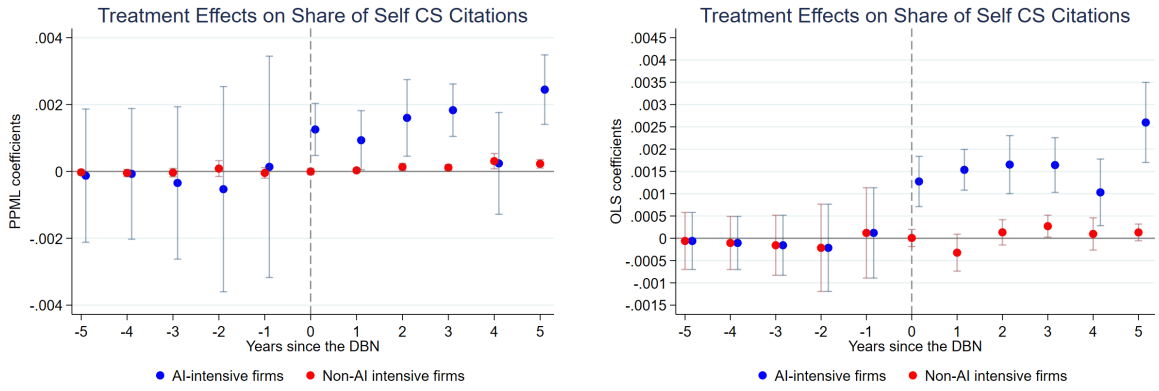


This figure illustrates the impact on the proportion of cybersecurity inventors within patent teams, calculated at the patent level and averaged annually at the firm level. Cybersecurity inventors are defined as those who have contributed to the development of at least one cybersecurity-related patent for the same assignee firm as the non-cybersecurity patent in question. Cybersecurity patents are identified through CPC classification codes as assigned by the USPTO, with further details provided in the main text.Top 500 firms are based on total sales in the year 2000, referred to as "superstar firms," adopting the terminology from Autor et al. (2020). The 'event year 0' corresponds to the staggered implementation of Data Breach Notification (DBN) laws across the United States. AI-intensive firms are identified based on criteria outlined earlier in the text.

More knowledge transfer. We go further and test whether AI-intensive firms exhibit more knowledge transfer between data security and other domains.

As shown in Figure (8), the share of *self*-data security patent citations for AI-intensive firms increases on impact, while the share of *self*-data security patent citations for non-AI-intensve firms stays flat. This suggests that AI-intensive firms cite their own data security (CS) patents much more often after the adoption of DBN laws.

Figure 8: Share of *self*-cybersecurity patent citations by AI intensity

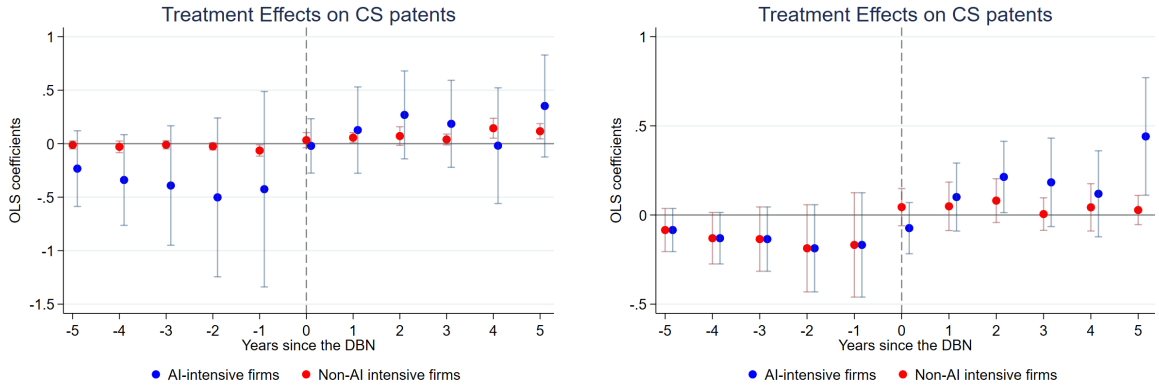


Legend: This figure plots the effects of the share of *self*-cybersecurity patent citations by firm AI-intensity pre- and post- treatment. The '0' event is the staggered adoption of DBN laws across the United States. Estimates for AI intensive firms are in blue, while estimates for non-AI intensive firms are in red. AI intensive firms are identified as described previously.

No increase in data-security innovation. Figure (9) presents the BJS-weighted dynamic heterogeneous treatment effects of citation-weighted *cybersecurity* patent counts by firm AI-intensity (AI). The left-hand panel allows heterogeneous pre-trends, while the right-hand panel assumes common pre-trends for both groups, but estimates ATT separately post-treatment.

As shown in Figure (9), AI-intensive firms exhibit a slight increase in cyber-security patenting, although the results are not significant except for long-term horizons. The insignificance of results is due to there being very few firms overall that produce cyber-security patents.

Figure 9: Citation-weighted *cybersecurity* patent count by AI intensity.

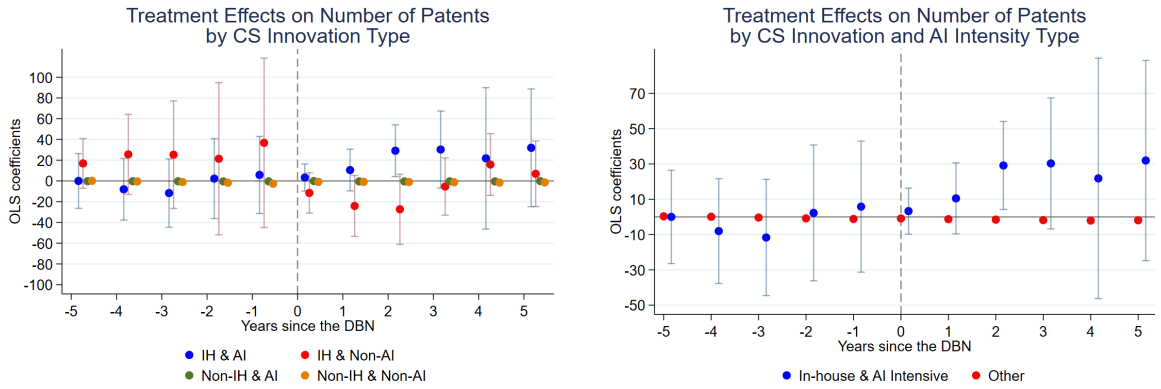


Legend: This figure plots the effects of citation-weighted cybersecurity patent counts by firm AI-intensity pre- and post- treatment. The '0' event is the staggered adoption of DBN laws across the United States. Estimates for AI intensive (DI) firms are in blue, while estimates for non-AI intensive (non-DI) firms are in red. AI intensive firms are identified as described previously.

Our analysis suggests that, while AI-intensive firms do not increase their overall issuance of data security (CS) patents, they increase their share of *self*-data security (CS) patent citations in their other patents after the adoption of DBN laws. This suggests a strong knowledge transfer from their data protection operations to their product and service development.

In-house data security firms. Lastly, we explore whether within the AI-intensive firm category, it is those firms which develop data security in-house, using their own inventors both for data security innovation and for non-data security innovation, that respond the most.

Figure 10: Citation-weighted patent count, data intensity interacted with in-house protection



Legend: This figure plots BJS-weighted dynamic heterogeneous treatment effects of citation-weighted cybersecurity patent counts by firm's choice of in-house vs. external cybersecurity protection interacted with data-intensity pre- and post- treatment. The '0' event is the staggered adoption of DBN laws across the United States. Estimates for in-house cybersecurity (in-house CS) firms are in blue, while estimates for non-in-house cybersecurity (non-in-house CS) firms are in red. In-house cybersecurity firms are identified if they cite at least one of their own cybersecurity patents in their general patents. Data intensity firms are identified as mentioned previously.

Figure (10) presents the BJS-weighted dynamic heterogeneous treatment effects of citation-weighted overall patent counts by firm AI-intensity (AI) interacted with in-house data security (in-house CS) protection choices. The left-hand panel allows heterogeneous pre-trends, while the right-hand panel assumes common pre-trends for both groups, but estimates ATT separately post-treatment.

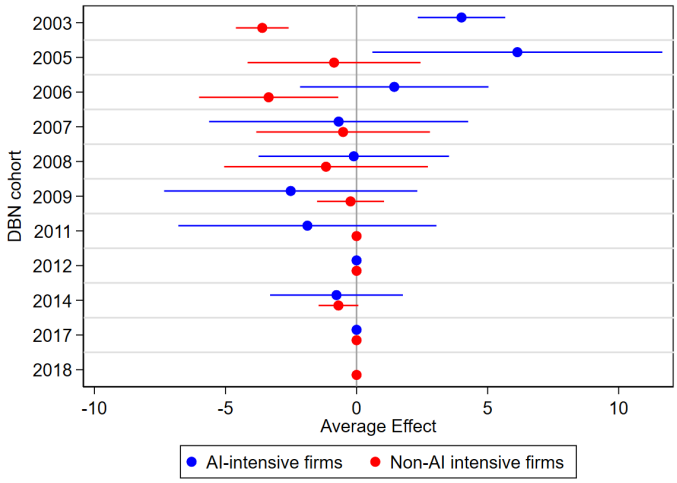
As shown in Figure (10), firms that use both in-house data protection and are AI-intensive exhibit a slight (temporary) increase in data-security patenting.

3.4 Cohort and calendar year effects

Cohort effects. We examine whether firms' innovation changes after the adoption of DBN laws depending on the cohort. This could happen if later treated cohorts

anticipate DBN law adoption in their state. In that case, the estimates will be smaller for later treated cohorts. It could also happen if the nature of data risk has changed over time in such a way that first movers had an advantage. Moreover, if data risk has changed in nature and severity over the last twenty years, it could be that it became too costly for later treated cohorts to invest resources into growth and innovation because too many resources had to go directly in managing data risk. While we cannot separate these mechanisms empirically, they could all be at play at the same time.

Figure 11: Treatment effects by DBN cohort



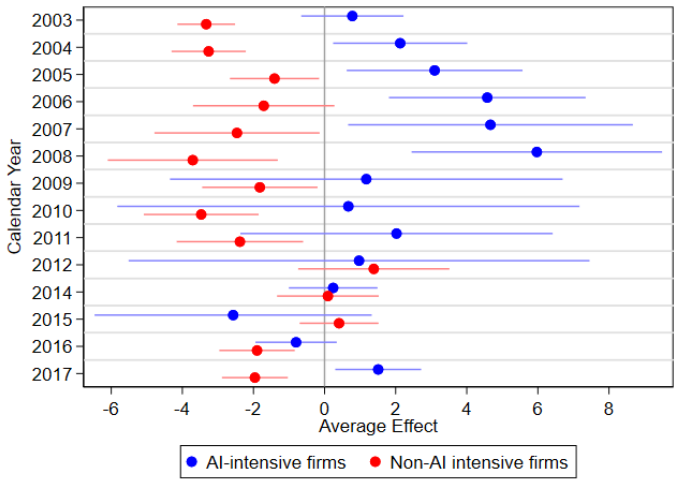
This figure plots treatment effects by cohort for AI-intensive and non-AI intensive firms. A cohort for a firm is the year in which the state it is incorporated in implements DBN laws. AI intensive firms are identified as mentioned previously. Estimates are average treatment in the first three years after DBN law implementation.

Figure (11) shows that, indeed, the very earliest treated AI-intensive cohorts responded the most. These effects are averaged across all the post-treatment years. We explain the effects being concentrated in the early part of the sample as due to the multi-state impact of these Data Breach Notification Laws. Firms operating across multiple states might experience the effects of regulatory changes before their headquarters state does. When a company is active in both regulated and non-regulated areas, policy effects could extend from the regulated to the non-regulated regions through the company's internal strategies, actions, or economic behavior, potentially influencing the comparison group. Such firms may alter their operational strategies across states in reaction to new regulations, including shifting resources to regions with more lenient regulations, which might influence the accuracy of assessing the regulation's effects. Nonetheless, these issues actually support our analysis, as they suggest that our estimates represent a minimum effect of Data Breach Notification (DBN) laws on regulated

firms compared to a comparison group that might have adapted in anticipation. Although no method is flawless, we are confident that our approach effectively accounts for corporate reactions to increased costs associated with data risk, by focusing on estimating minimum effects.

Calendar year effects. In Figure (12) we explore the DBN law effects in a particular year. The period 2004 to 2008 is the most intense period in terms of increase in firms' innovation activities in response to the increase in data risk, as measured by DBN laws.

Figure 12: Treatment effects by calendar year



Legend: This figure plots treatment effects by calendar year for AI-intensive and non-AI intensive firms. Each estimate is the sum of effects of all treated cohorts up to and including that year in that particular year. So for instance 2008 will contain the effect for 3rd year since the firms treated in 2005, 2nd year effects for firms treated in 2006 and so on. Treatment effects up to 5 years contribute to the calculations. AI intensive firms are identified as mentioned previously.

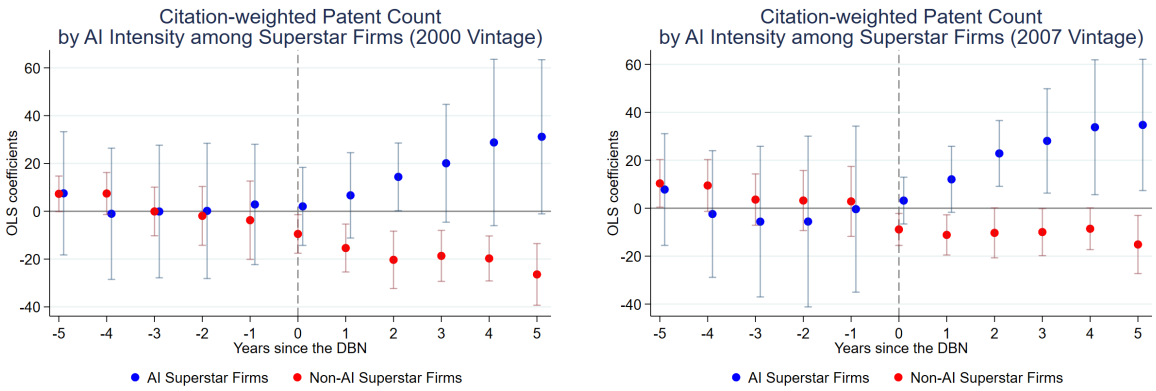
Similar to the cohort effects, companies across multiple states might feel policy impacts earlier, influencing their operations and possibly affecting control group comparisons. Adjustments in response to regulations can skew the true effect estimation to the earlier sample, as seen in Figure (12).

3.5 Sub-sample analysis

Superstar firms. We also examine whether the effects are concentrated in superstar firms and we find that it is not size driving the results, as shown in Appendix B. Superstar firms, characterized by significant market share, high profitability, and often global reach, have increasingly leveraged advanced technologies, including artificial intelligence (AI), to maintain and enhance their market positions. These firms often have

the resources and strategic motivations to invest in AI-intensive processes. However, as shown in Figure 13, within the universe of superstar firms, it is the AI-intensive firms that display a positive reduced-form correlation between data risk and firm outcomes. Non-AI-intensive superstar firms experience negative effects.

Figure 13: Citation-weighted patent count within superstar firms: by AI intensity

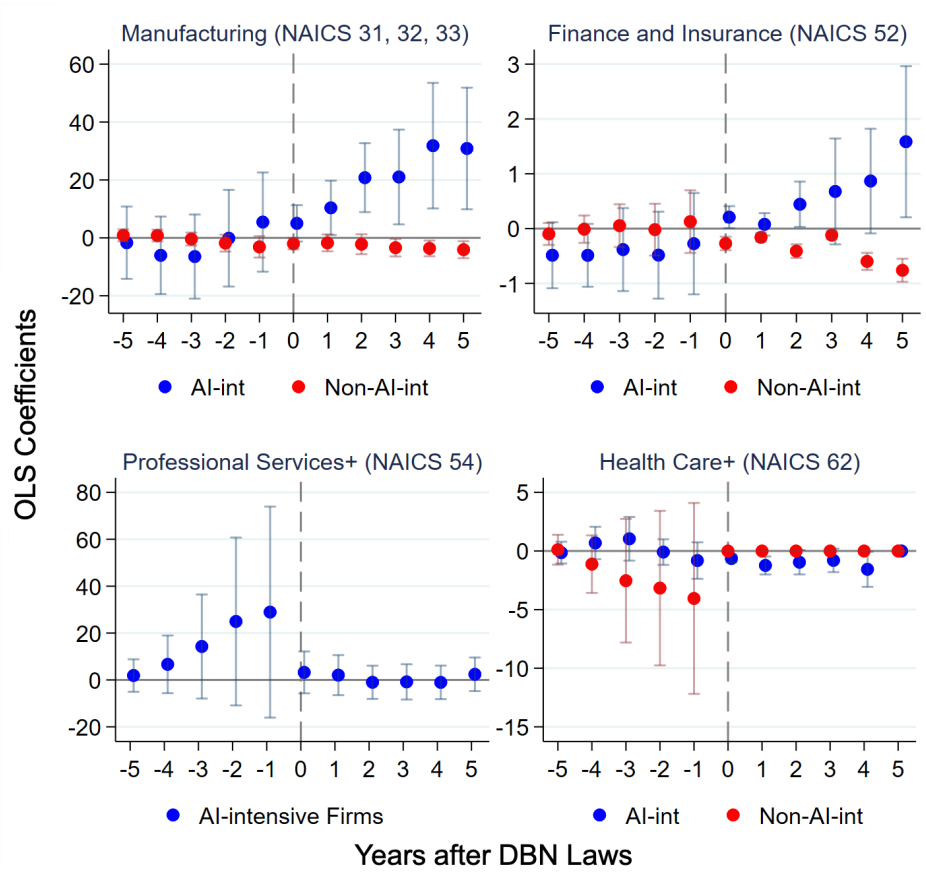


This figure plots BJS dynamic heterogeneous treatment effects on citation-weighted patent counts by firm AI-intensity pre- and post- treatment. The '0' event is the staggered adoption of DBN laws across the United States. AI intensive firms are identified using a combination of the USPTO dataset on AI patents (Giczy et al. (2022)) and the KPSS patent dataset linked to firms (Kogan et al. (2017)). Moreover, firms that are close to AI patenting firms in the sense of Hoberg and Phillips (2016) and mirror their AI innovations are also considered AI-intensive. This measure has the advantage to be constructed from entirely publicly available data and it is different from IT expenditures. Analysis conducted on the highest 500 firms in terms of total sales for a given vintage year. Where vintage year could be 2000 or 2007. These firms are referred to as 'superstar firms,' following the terminology used by Autor et al. (2020).

Finance and tech firms. Lastly, we examine, within the universe of AI-intensive firms, which industries exhibit the strongest effects. Interacting our lagged data-risk score measure with an industry dummy for each of the NAICS industry classification categories, we observe an increase in innovation activities and profitability for AI-intensive firms that offer financial and tech products (belonging either to 'Finance and Insurance', or to 'Manufacturing', as shown in Figure 14.

In Appendix B, we show the results for all the NAICS industry classifications. We observe strong positive effects for firms in 'Manufacturing', 'Retail', and 'Finance and Insurance', and negative effects for firms in 'Agriculture', 'Health care' and 'Accommodation', with no significant results for other industry classifications. Upon closer inspection, firms in 'Manufacturing' and 'Retail' are high-tech firms such as Amazon and Apple, whose firm boundaries have been expanding rapidly in the last 20 years. They not only offer retail products, but also tech and financial products.

Figure 14: Citation-weighted patent count: by AI intensity in select industries



This figure plots BJS dynamic heterogeneous treatment effects of citation-weighted patent counts. Industry refers to North American Industry Classification System (NAICS) at the two-digit level. The '0' event is the staggered adoption of DBN laws across the United States. Estimates for AI-intensive firms. AI intensive firms are identified using a combination of the USPTO dataset on AI patents (Giczy et al. (2022)) and the KPSS patent dataset linked to firms (Kogan et al. (2017)). Moreover, firms that are close to AI patenting firms in the sense of Hoberg and Phillips (2016) and mirror their AI innovations are also considered AI-intensive. This measure has the advantage to be constructed from entirely publicly available data and it is different from IT expenditures. The figure plots the estimates only for those industries where a separate estimation was possible for the AI intensive firms.

The differential response to increased data risk across industries, particularly between the financial, insurance, tech sectors, and the health and accommodation sectors, reflects a complex interplay of factors. Financial and insurance industries, alongside tech firms, are fundamentally data-centric and technology-driven. For these sectors, heightened data risk prompts strategic investments in advanced data security measures and data management technologies, transforming potential vulnerabilities into catalysts for innovation, efficiency, and market differentiation. These industries are also buoyed by substantial financial resources and a regulatory environment that, while demanding, incentivizes robust data protection practices, making them more resilient and adaptive to data-related challenges.

In stark contrast, the health and accommodation sectors face more daunting ob-

stacles when confronting increased data risks. Stringent regulations, such as HIPAA in healthcare, impose rigorous data privacy obligations, turning enhanced data risks into significant compliance burdens. These industries, less focused on data management and more on service delivery, often lack the technological infrastructure and expertise to navigate the evolving landscape of data threats effectively. The result is a pronounced vulnerability to operational disruptions and financial liabilities arising from data breaches, without the offsetting advantages of technological innovation or new revenue streams that tech-savvy sectors enjoy. This divergence underscores the critical need for industry-specific strategies in addressing and mitigating data risks.

4 A model of big data, data risk and data security

4.1 Efficient data use and security risks

We consider a competitive industry. Time is discrete and infinite. There is a continuum of firms indexed by i. Each firm i produces a good of quality $A_{i,t}$.

$$y_{i,t} = A_{i,t}. \tag{3}$$

Because the single input employed in production is one unit of capital, variable $A_{i,t}$ also represents the real value of the producer's output.

Quality $A_{i,t}$ depends on a firm's choice of a production technique $a_{i,t}$, which can be interpreted as managing inventories, or learning about consumers' tastes. In each period, and for each firm, there is one optimal technique with a persistent and a transitory components: $\theta_{i,t} + \epsilon_{a,i,t}$. The persistent component $\theta_{i,t}$ is unknown and follows an AR(1) process, where $\eta_{i,t}$ is $i.i.d.$ across time and firms:

$$\theta_{i,t} = \bar{\theta} + \rho(\theta_{i,t-1} - \bar{\theta}) + \eta_{i,t}. \tag{4}$$

Firms have a noisy prior about the realization of θ_0. The transitory shock $\epsilon_{a,i,t}$ is $i.i.d.$ across time and firms and is unlearnable. Deviating from that optimum incurs a quadratic loss in quality:

$$A_{i,t} = \bar{A}_i - (a_{i,t} - \theta_{i,t} - \epsilon_{a,i,t})^2. \tag{5}$$

Quality $A_{i,t}$ is a strictly decreasing function of the difference between the firm's chosen production technique, $a_{i,t}$, and the optimal technique $\theta_{i,t} + \epsilon_{a,i,t}$. A decreasing function means that techniques far away from the optimum result in worse quality goods.

Data as by-product. Data helps firms infer $\theta_{i,t}$ Term ϵ_a indicates that firms are incapable of fully inferring $\theta_{i,t}$ at the end of each period; it makes the accumulation of past data a valuable asset. If a firm knew the current value of $\theta_{i,t}$, it would maximize quality by setting $a_{i,t} = \theta_{i,t}$.

In our model, similar to Farboodi et al. (2019) and Farboodi and Veldkamp (2021), data is a by-product of economic activity. Each firm passively obtains z data points as a by-product of production. Each data point $m \in [1 : z]$ reveals

$$s_{i,t,m} = \theta_{i,t} + \epsilon_{i,t,m}, \tag{6}$$

where $\epsilon_{i,t,m}$ is *i.i.d.* across firms, time, and signals. For tractability, we assume that all the shocks are normally distributed: fundamental uncertainty is $\eta_{i,t} \sim N(\mu, \sigma_\theta^2)$, signal noise is $\epsilon_{i,t,m} \sim N(0, \sigma_\epsilon^2)$, and the unlearnable quality shock is $\epsilon_{a,i,t} \sim N(0, \sigma_a^2)$.

data risk. Data is subject to data risk or data incident risk, meaning that it can be lost and, in that case, it can no longer be used for prediction. We denote the degree of data risk by $\vartheta \in [0, 1]$. With probability ϑ, a firm risks losing all its data, while with probability $(1 - \vartheta)$ the firm keeps its data generated as a by-product of activity, $z\sigma_\theta^2$. Thus, the data endowment under data risk is $(1 - \vartheta)z\sigma_\epsilon^2$.

Data security. A key assumption of our model is that firms are heterogeneous in their capability to protect themselves against data risk. High capability (H-type) firms can develop in-house data security protection, while low capability (L-type) firms cannot develop this security internally, but can buy it externally from H-type firms.

The essential distinction between in-house and external data security is that internal data security can also be used to innovate, besides providing protection against data loss and destruction. This is because in-house data security is typically more easily integrated with existing R&D and product development systems, and tends to be more tailored for a firms' specific business needs. In the model, innovation is modeled as an increase in the productivity ceiling $\overline{A}_i$.

Low-type capability firms do not generate in-house security, but they can buy it externally from High H-type firms. In this case, they can only use it to mitigate the impact of data risk and not to innovate (i.e., they can use the security software for protecting their production process, but their R&D department does not know and is unable to use the security software for product improvements).

Let m_H represent the share of H-type firms. Aggregate output is then the sum of

weighted outputs for the two types of firms:

$$Y_t = \int_0^1 A_{i,t} di = m_H A_{H,t} + (1 - m_H) A_{L,t}. \tag{7}$$

Let $\tau_t \geq 0$ represent the investment in in-house data security made by a firm of the H-type. Let also $\delta_t \geq 0$ represent the amount of external data security bought by a firm of the L-type from the H-type firms at an endogenous price denoted by π. Given the firms' shares, the amount of protection that is sold by a H-type producer must be $\frac{1-m_H}{m_H}\delta_t$. In this case, on the aggregate H-firms sell $(1 - m_H)\delta_t$, which is precisely the value of protection purchased by the universe of L-type firms.

Nonrivlary. When a company invests in data security measures such as firewalls, encryption protocols, or security software, these measures protect the company's data and systems without necessarily reducing their effectiveness for other companies that may use similar security tools. This suggests that data security is (partially) nonrival.

Thus, we assume that when an H-type firm sells a given amount of cyber protection, it retains, for its own use, a share $1 - \iota$ of such protection, where $\iota \in (0, 1)$. Therefore, the H-firm that invests τ_t in data security and trades $\frac{1-m_H}{m_H}\delta_t \leq \tau_t$, will retain, for its own use, $\tau_t - \iota\frac{1-m_H}{m_H}\delta_t$. This amount of cyber protection can be used to mitigate the impact of data risk, transforming the term $(1 - \vartheta)z$ into $\left[1 - \vartheta e^{-\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}\right] z$. Note that if $\tau_t - \iota\frac{1-m_H}{m_H}\delta_t = 0$, there is no use of cyber protection, and the effect of data risk over data is maximum; if $\tau_t - \iota\frac{1-m_H}{m_H}\delta_t \to \infty$, then there is full protection, and the original data endowment maintains its integrity.

Firm problem. With this in mind, we can write firm i's optimization problem, where $i \in \{H, L\}$. As mentioned previously, the H-type firm can use the investment in data security to enhance the potential quality of the produced good. Hence, constant $\overline{A}_i$ is replaced, for this type of firm, by the term $\overline{A}e^{b\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}$.

An H-type firm chooses a sequence of quality decisions $a_{i,t}$, in-house data security investments τ_t, and how much data security δ_t to sell at price π_t to maximize:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\overline{A}e^{b\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)} - (a_{i,t} - \theta_{i,t} - \epsilon_{a,i,t})^2 - \tau_t + \frac{1 - m_H}{m_H}\delta_t\pi_t - r\right] \tag{8}$$

An L-type firm chooses a sequence of quality decisions $a_{i,t}$, and how much external

data protection δ_t to buy at price π_t to maximize:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\bar{A} - (a_{i,t} - \theta_{i,t} - \epsilon_{a,i,t})^2 - \delta_t \pi_t - r \right] \tag{9}$$

Note the differences between the two expressions: innovation from data security is possible for the H-type firm but not for the L-type firm ; the cost of investment in data security is present only in the H-type firm expression; protection trading is a revenue for those who sell it and a cost for those who buy it.

The stock of knowledge. The information set of firm $i \in \{H, L\}$ when it chooses its technique $a_{i,t}$ is $\mathcal{I}_{i,t} = \{\mathcal{I}_{i,t-1}, \{s_{i,t-1,m}\}_{m=1}^{z}, A_{i,t-1}\}$ where z is the net numbers of points added each period as a by-product of economic activity. To make the problem recursive, we construct a helpful summary statistic for this information, called the "stock of knowledge." A firm's stock of knowledge is the inverse of its posterior variance, or in other words, the precision of firm i's forecast of θ_t, which is formally:

$$\Omega_{i,t} = \mathbb{E}\left[(\mathbb{E}[\theta_t | \mathcal{I}_{i,t}] - \theta_t)^2 \right]^{-1} \tag{10}$$

Note that the inside of the expression is the difference between a forecast, $\mathbb{E}[\theta_t|\mathcal{I}_{i,t}]$ and the realized value, θ_t, and is therefore a forecast error. An expected squared forecast error is the variance of the forecast. It is also called the variance of θ_t, conditional on the information set $\mathcal{I}_{i,t}$, or the posterior variance. The inverse of a variance is a precision. Thus, this is the precision of firm i's forecast of θ_t.

4.2 A law of motion for knowledge

The state variables of the recursive problems in (8) and (9) are the prior mean and variance of beliefs about $\theta_{i,t-1}$, and the new data points. Taking a first order condition with respect to the technique choice, we find that the optimal technique is $a^*_{i,t} = \mathbb{E}_i[\theta_{i,t}|\mathcal{I}_{i,t}]$. Given the posterior variance of beliefs in equation (10), the expected quality for the H-type and the L-type firms, respectively, are

$$\mathbb{E}[A_{H,t}] = \overline{A} e^{b\left(\tau_t - \iota \frac{1-m_H}{m_H} \delta_t\right)} - \Omega_{H,t}^{-1} - \sigma_a^2 \tag{11}$$

$$\mathbb{E}[A_{L,t}] = \overline{A} - \Omega_{L,t}^{-1} - \sigma_a^2 \tag{12}$$

Deriving the law of motion for the stock of knowledge, $\Omega_{i,t}$, requires adding new data from two sources: 1) data as a by-product of production, which is subject to cyberrisk but can be protected through data security and 2) data inferred from a firm observing

its own quality at the end of a production period. These two pieces of information are incorporated into beliefs using Bayes' law.

Each firm $i \in \{H, L\}$ observes $z_i = z$ data points as a by-product of economic activity. This means that the sum of the precisions of all the signals (data proints), $z_i \sigma_\epsilon^{-2}$ is part of the stock of knowledge. Both types of firms, the H-type and the L-type, are subject to cyberrisk, which can be reduced through protection. The H-type firm reduces cyberrisk by the amount of cybersecurity it retains for its own use, $\tau_t - \iota \frac{1-m_H}{m_H} \delta_t \leq \tau_t$, after it invests τ_t in cybersecurity and trades $\frac{1-m_H}{m_H} \delta_t \leq \tau_t$ cyber protection which is non-rival. This amount of cyberprotection can be used to mitigate the impact of cyberrisk, implying that the weighted sum of precisions of data points obtained as a byproduct of economic activity, subject to cyberrisk and after optimal cybersecurity decisions, is $\left[1 - \vartheta e^{-\left(\tau_t - \iota \frac{1-m_H}{m_H} \delta_t\right)}\right] z \sigma_\epsilon^{-2}$. L-type firm buys protection in amount δ_t and, therefore, the weighted sum of precisions of data points obtained as a byproduct of economic activity, subject to cyberrisk and after optimal cybersecurity decisions, is $\left[1 - \vartheta e^{-\delta_t}\right] z \sigma_\epsilon^{-2}$.

Moreover, each firm $i \in \{H, L\}$ is also learning from seeing its own realization of quality $A_{i,t}$ at the end of each period t, with precision σ_a^{-2}. This information is different from the produced data because the quality realization is a signal about θ_t, not about θ_{t+1}. Therefore, σ_a^{-2} gets added to the time-t stock of knowledge and depreciates, just like other time-t knowledge that the firm takes with it to time $t+1$.

Lemma (1) expresses the dynamic knowledge constraint that puts together data depreciation and data inflows.

Lemma 1 *The dynamic knowledge constraint is, for the H-type firm:*

$$\Omega_{H,t+1} = \left[\rho^2 (\Omega_{H,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \left[1 - \vartheta e^{-\left(\tau_t - \iota \frac{1-m_H}{m_H} \delta_t\right)}\right] z \sigma_\epsilon^{-2} \tag{13}$$

The L-type firm buys protection in amount δ_t and, therefore,

$$\Omega_{L,t+1} = \left[\rho^2 (\Omega_{L,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \left(1 - \vartheta e^{-\delta_t}\right) z \sigma_\epsilon^{-2} \tag{14}$$

In this last case, if the firm buys no protection, data loss risk occurs in a share ϑ; if it buys infinite protection, it faces no data risk.

The demonstration for this lemma and all subsequent lemmas and propositions can be found in the Appendix. The proof involves utilizing Bayes' law, or alternatively, the Ricatti equation within a modified Kalman filter framework. Given the similarity in information structure to that of a Kalman filter, the sequence of conditional variances (or conversely, their inverses, the sequence of precisions) is deterministic.

4.3 Recursive representation of the firm's problem, equilibrium and steady state

Lemma (2) proceeds with the recursive representation of the expected firm value.

Lemma 2 *The optimal sequences of in-house data security investments* $\{\tau_t\}$ *and data security sales* $\{\delta_t\}$ *solve the following current-value Hamiltonian function for the H-type firm:*

$$H_{H,t}(\Omega_{H,t},\tau_t,\delta_t,p_{H,t}) = \overline{A}e^{b\left(\tau_t-\iota\frac{1-m_H}{m_H}\delta_t\right)} - \Omega_{H,t}^{-1} - \sigma_a^2 - \tau_t + \frac{1-m_H}{m_H}\delta_t\pi_t - r + \tag{15}$$
$$+ \beta p_{H,t+1}(\Omega_{H,t+1}-\Omega_{H,t})$$
$$\text{where}\quad \Omega_{H,t+1} = \left[\rho^2(\Omega_{H,t}+\sigma_a^{-2})^{-1}+\sigma_\theta^2\right]^{-1} + \left[1-\vartheta e^{-\left(\tau_t-\iota\frac{1-m_H}{m_H}\delta_t\right)}\right] z\sigma_\epsilon^{-2} \tag{16}$$

and $p_{H,t}$ *is the shadow-price or co-state variable associated with the state variable, and the transversality condition is* $\lim_{t\to\infty}\Omega_{H,t}\beta^t p_{H,t}=0$.

The optimal sequence of cybersecurity purchases $\{\delta_t\}$ *solve the following current-value Hamiltonian function for the L-type firm:*

$$H_{L,t}(\Omega_{L,t},\tau_t,\delta_t,p_{L,t}) = \overline{A} - \Omega_{L,t}^{-1} - \sigma_a^2 - \delta_t\pi_t - r + \beta p_{L,t+1}(\Omega_{L,t+1}-\Omega_{L,t}) \tag{17}$$
$$\text{where}\quad \Omega_{L,t+1} = \left[\rho^2(\Omega_{L,t}+\sigma_a^{-2})^{-1}+\sigma_\theta^2\right]^{-1} + \left(1-\vartheta e^{-\delta_t}\right) z\sigma_\epsilon^{-2} \tag{18}$$

and $p_{L,t}$ *is the shadow-price or co-state variable associated with the state variable, and the transversality condition is* $\lim_{t\to\infty}\Omega_{L,t}\beta^t p_{L,t}=0$.

See the Appendix for the proof. This result greatly simplifies the problem by collapsing it to a deterministic dynamic system involving only one state variable, $\Omega_{i,t}$, where $i=H$ or $i=L$. The reason we can do this is that quality $A_{i,t}$ depends on the conditional variance of $\theta_{i,t}$ and because the information structure is similar to that of a Kalman filter, where the sequence of conditional variances is generally deterministic.[5] This Kalman system has a 2-by-1 observation equation, with $n_{i,t}=z$ signals about $\theta_{i,t}$ and one signal about $\theta_{i,t-1}$. The signal about $\theta_{i,t-1}$ comes from observing last period's output, which reveals quality $A_{i,t-1}$, which, in turn, reveals $\theta_{i,t}+\epsilon_{a,i,t}$.[6]

[5]The optimal choice of technique is always the same: $a_{i,t}^* = \mathbb{E}_i[\theta_{i,t}|\mathcal{I}_{i,t}]$. The way $a_{i,t}$ enters into expected quality $A_{i,t}$ is through $\mathbb{E}[(\mathbb{E}[\theta_{i,t}|\mathcal{I}_{i,t}]-\theta_{i,t})^2]$, which is the conditional variance $\Omega_{i,t}$. We can replace the entire sequence of $a_{i,t}^*$ with the sequence of variances, which is deterministic here because of normality. The only randomness in this model comes from the signals and their realizations, but they never affect the conditional variance, since normal means and variances are independent. Thus, given $\Omega_{i,t-1}$, $\Omega_{i,t}$ is a sufficient statistic for $n_{i,t}=z$ and $\Omega_{i,t+1}$. The mean $\mathbb{E}[\theta_{i,t}|\mathcal{I}_{i,t}]$ is not a state variable because it only matters for determining $a_{i,t}$ and does not affect anything else.

[6]Firms observe $(\theta_{i,t}+\epsilon_{a,i,t})^2$. For tractability, we assume that firms know whether the root is

Equilibrium. From the Hamiltonian functions, and assuming all variances are equal such that $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2$, we can derive the equilibrium conditions.

$$\frac{\partial H_{H,t}}{\partial \tau_t} = 0 \Rightarrow \beta p_{H,t+1} = \frac{1 - b\overline{A}e^{b\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}}{\vartheta e^{-\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)} z\sigma^{-2}} \tag{19}$$

$$\frac{\partial H}{\partial \delta_t} = 0 \Rightarrow \beta p_{H,t+1} = \frac{\pi_t - b\overline{A}\iota e^{b\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}}{\vartheta\iota e^{-\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)} z\sigma^{-2}} \tag{20}$$

$$\beta p_{H,t+1} - p_{H,t} = -\frac{\partial H}{\partial \Omega_{H,t}} \Rightarrow \left[\rho + \frac{\sigma^2}{\rho}(\Omega_{H,t} + \sigma^{-2})\right]^{-2} \beta p_{H,t+1} = p_{H,t} - \Omega_{H,t}^{-2} \tag{21}$$

From (32) and (33), it emerges a constant optimal trading price, which is simply $\pi_t = \iota$. The price of protection is directly associated with the degree of its own nonrivalry. If protection is completely non-rival (i.e., $\iota = 0$), then its price is zero; if protection is fully rival, its price is 1.

For the L-type firm, the equilibrium conditions are:

$$\frac{\partial H}{\partial \delta_t} = 0 \Rightarrow \beta p_{L,t+1} = \frac{\pi_t}{\vartheta e^{-\delta_t} z\sigma^{-2}} \tag{22}$$

$$\beta p_{L,t+1} - p_{L,t} = -\frac{\partial H}{\partial \Omega_{L,t}} \Rightarrow \left[\rho + \frac{\sigma^2}{\rho}(\Omega_{L,t} + \sigma^{-2})\right]^{-2} \beta p_{L,t+1} = p_{L,t} - \Omega_{L,t}^{-2} \tag{23}$$

Steady-state. The steady-state of the economy is characterized by a level of data security held by H-type firms after trade given by:

$$\tau^* - \iota\frac{1 - m_H}{m_H}\delta^* = -\ln\left(\frac{z - \Xi_H}{\vartheta z}\right) \tag{24}$$

where $\Xi_H \equiv \left\{\Omega_H^* - [\rho^2(\Omega_H^* + \sigma^{-2})^{-1} + \sigma^2]^{-1}\right\}\sigma^2$. At steady-state, the amount of protection bought by L-type firms is given by:

$$\delta^* = -\ln\left(\frac{z - \Xi_L}{\vartheta z}\right) \tag{25}$$

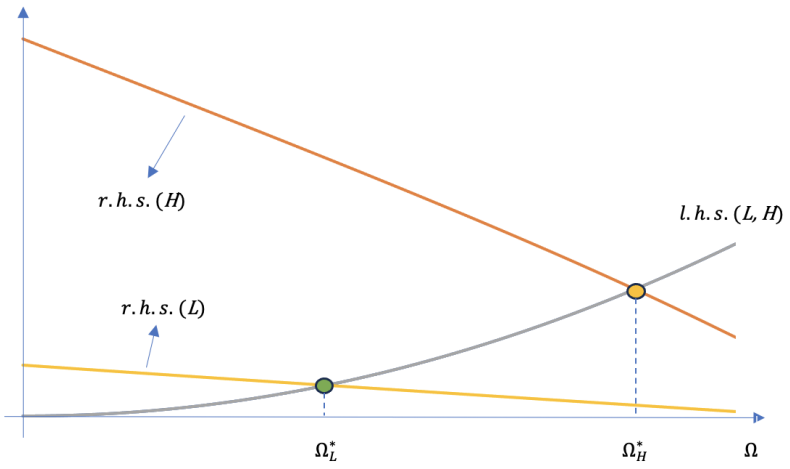
with $\Xi_L \equiv \left\{\Omega_L^* - [\rho^2(\Omega_L^* + \sigma^{-2})^{-1} + \sigma^2]^{-1}\right\}\sigma^2$.

Figure (15) plots the equilibrium knowledge levels of this economy. The demand

positive or negative. For more on this and for the derivation of the belief updating equations, see online Appendix.

and supply of knowledge for H-type firms intersect at a higher level than the demand and supply of knowledge for L-type firms. The demand of L-type firms is flatter and more inelastic than the demand of H-type firms. Thus, in equlibrium, H-type firms end up with a higher level of knowledge than L-type firms.

Figure 15: Steady-state stocks of knowledge.



Legend: The figure shows the equilibria levels of knowledge for H-type firms (in orange on the right) and L-type firms (in green on the left) as a function of the cyberrisk index, ϑ, on the X-axis. H-type firms achieve a higher level of steady-state knowledge than L-type firms. The parameters used in this simulations are the following: $z = 10$, $\rho = 0.9$, $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, $m_H = 1/3$, $\iota = 0.6$, $\beta = 0.96$, $\vartheta = 0.75$, $\overline{A} = 25$, $b = 0.035$, and $r = 1$.

Table (5) illustrates the steady state equilibrium of this economy. In the case of this example, in the steady state, H-type firms invest 1.296 in in-house cyber protection, sell 0.130 cyber protection to L-type firms and remain with a cyber protection level of 0.335, which is higher than the L-type's level of protection of 0.130. In steady-state, knowledge, quality and profits are all higher for the H-type firm than for the L-type firm.

Table 5: Steady-state.

Parameter	**Symbol**	**Steady-state**
Knowledge H-type	$\Omega_H^\star$	3.224
Knowledge L-type	$\Omega_L^\star$	1.609
In-house cyberprotection	$\tau^\star$	1.296
Datasecurity traded	$\delta^\star$	0.130
Quality H-type	A_H^*	23.207
Quality L-type	A_L^*	21.879
Profits H-type	Π_H^*	21.068
Profits L-type	Π_L^*	20.800
Total output	Y	22.321

Legend: The parameters used in this simulations are the following: $z = 10$, $\rho = 0.9$, $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, $m_H = 1/3$, $\iota = 0.6$, $\beta = 0.96$, $\vartheta = 0.75$, $\overline{A} = 25$, $b = 0.035$, and $r = 1$.

4.4 Results and implications

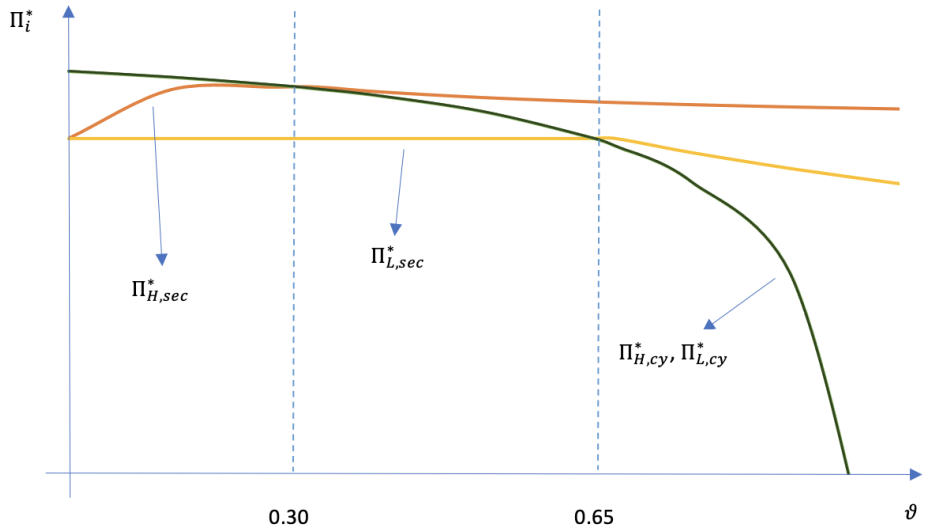
Throughout this section, a numerical example is employed with the goal of highlighting some of the most meaningful results of the model. These results comprise the impact of data risk over firms' profits, the timing of the decisions to engage in in-house cyber protection and external purchase of data security, and aggregate output.

4.4.1 Data protection helps firms hedge data risk

Our first numerical experiment studies how an increase in firm data risk changes firms' profits. We start by simulating firm profits in a model with no cyber protection. Then, we turn on data security protection for both types of firms to observe how their profits change. To compute the change in firms profitability when they face increasingly higher data risk, we change the data risk index ϑ continuously from no data risk ($\vartheta = 0$) to maximum data risk ($\vartheta = 1$) and re-compute the steady state. Figure (16) shows that the profits of H-type firms with data security fall by less than the profits of L-type firms as data risk increases. Moreover, the profits of both types of firms without data security protection at all drop dramatically as the overall level of data risk increases in the economy.

Without protection, the profits (in green) of H-type firms are the same as the profits of L-type firms and decreasing in the data risk index ϑ. Initially, the profits without data security decline slowly, but after the second threshold, they decline rapidly because the cost incurred in knowledge loss increases exponentially with data risk without protection. With protection, however, the profits of H-type firms (in orange) are always higher than the profits of L-type firms (in yellow). And, as data risk increases,

Figure 16: Profits as a function of data risk.



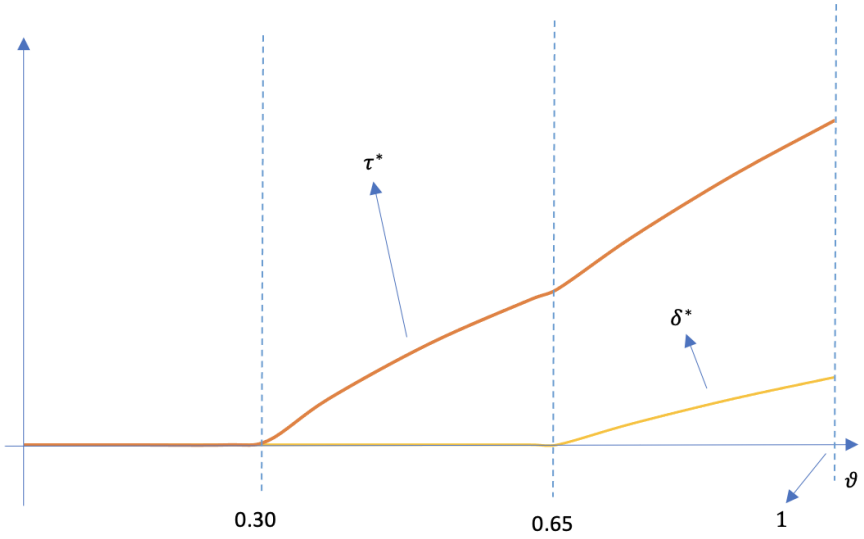
Legend: This figure plots the steady-state profit levels for H-type firms with (in orange, $\Pi^{\star}_{H,sec}$) and without cyberprotection (in green, $\Pi^{\star}_{H,cy}$), and L-type firms with (in yellow, $\Pi^{\star}_{L,sec}$) and without cyber protection (in green, $\Pi^{\star}_{L,cy}$), as a function of the data risk index, ϑ, on the X-axis. The parameters used in this simulation are the following: the data endowment $z = 10$, the coefficient of the AR(1) process $\rho = 0.9$, all variances $\sigma^2_\theta = \sigma^2_a = \sigma^2_\epsilon = \sigma^2 = 2.5$, the share of H-type firms $m_H = 1/3$, the non-rivalry parameter $\iota = 0.6$, the intertemporal discount factor $\beta = 0.96$, the data risk index $\vartheta = 0.75$, the maximum quality threshold $\overline{A} = 25$, the innovation externality $b = 0.035$, and the cost of capital $r = 1$.

the profits H-type firms decrease at a smaller rate than the profits of L-type firms (in yellow). An interesting observation is that initially, with protection, the profits of H-type firms first increase because the benefit of protection (which is data security-driven innovation) is initially higher than the cost of cyber crime.

4.4.2 High capability firms engage in protection at lower risk levels than L-type firms

What governs the steady-state size of firms is firms' cyber protection levels as a function of the cyberrisk index, ϑ, plotted in Figure (17).

Figure 17: Dataprotection as a function of cyberrisk.



Legend: The figure plots in-house data security investment, τ_t, by H-type firms (in orange), and external data security acquisition by L-type firms (in yellow). Notice the two critical thresholds at which in-house cyber protection and external cyber protection become strictly positive. The parameters used in this simulations are the following: the data endowment $z = 10$, the coefficient of the AR(1) process $\rho = 0.9$, all variances $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, the share of H-type firms $m_H = 1/3$, the non-rivalry parameter $\iota = 0.6$, the intertemporal discount factor $\beta = 0.96$, the data risk index $\vartheta = 0.75$, the maximum quality threshold $\overline{A} = 25$, the innovation externality $b = 0.035$, and the cost of capital $r = 1$.

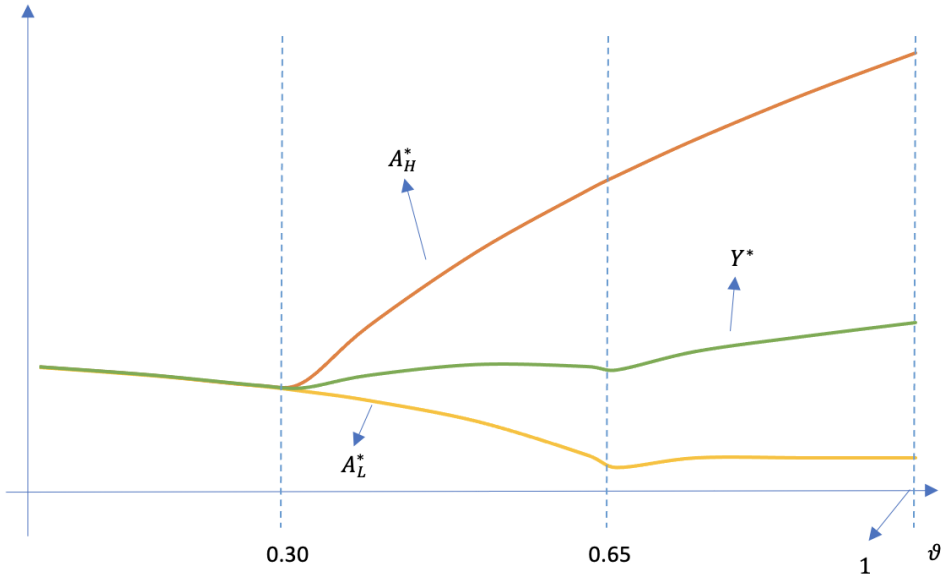
Evaluating the model for different values of ϑ, and letting all other parameters be as before, we find two critical thresholds: at $\vartheta = 0.6583$, optimal data security purchases, δ^*, changes from negative to positive, implying that L-type firms buy protection only for $\vartheta > 0.6583$. For $\vartheta \leq 0.6583$, H-type firms have to choose whether to invest in protection or not, knowing that they cannot sell any cyber protection. H-type firms are indifferent between investing in protection or not at a critical threhold level of $\vartheta = 0.3$. For $\vartheta > 0.3$, H-type firms invest in protection, otherwise they do not.

4.4.3 Data risk can also sustain growth

Surprisingly, while one expects aggregate economic output to be decreasing in data risk, there is a counteracting force that works especially at high levels of risk. This is shown in Figure (18).

Firms with a high capacity for in-house data security protection (in orange) use this protection to innovate, which raises the quality and quantity of production. Indeed, output is increasing in data risk for H-type firms at moderate to high levels of data risk. L-type firms do not have this positive spillover, because they only use data security for their own protection, to mitigate the negative effects of data risk. The aggregate output is a weighted average of the output of the two types of firms. Concerning the evolution of Y^* as ϑ increases, one notices that an initial fall is counteracted when H-type firms start to invest in protection, and this process gains momentum when L-type firms start protecting as well.

Figure 18: Output as a function of cyberrisk.

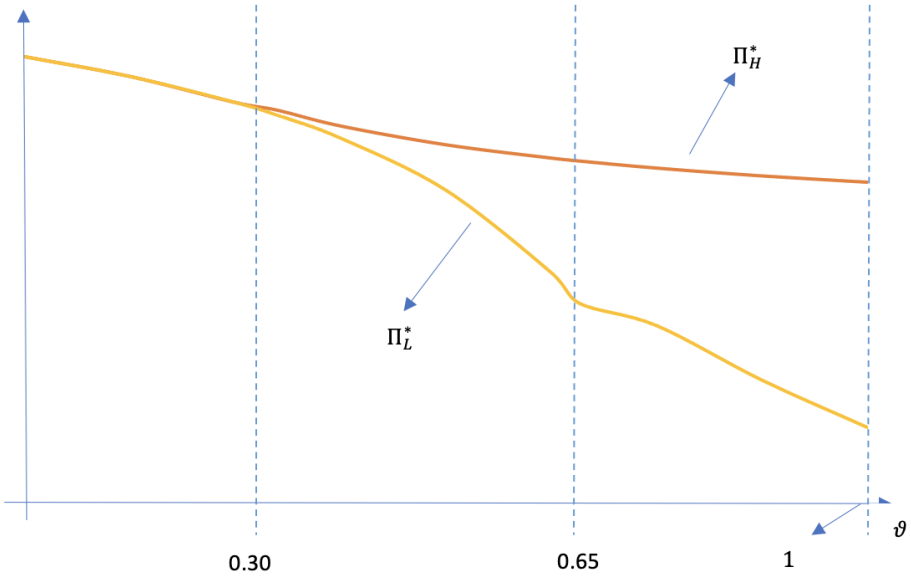


Legend: The parameters used in this simulations are the following: the data endowment $z = 10$, the coefficent of the AR(1) process $\rho = 0.9$, all variances $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, the share of H-type firms $m_H = 1/3$, the non-rivalry parameter $\iota = 0.6$, the intertemporal discount factor $\beta = 0.96$, the cyberrisk index $\vartheta = 0.75$, the maximum quality threshold $\overline{A} = 25$, the innovation externality $b = 0.035$, and the cost of capital $r = 1$.

We can recover the representation of profits in Figure (16), to plot the actual profits, given the choices of firms on whether to get protection or not. Figure (19) clarifies again the existence of three stages and the fact that data risk is much less harmful for H-type firms, because these make use of the innovation externality that

cybersecurity allows for.

Figure 19: Realized (equilibrium) profits.



Legend: The parameters used in this simulations are the following: the data endowment $z = 10$, the coefficent of the AR(1) process $\rho = 0.9$, all variances $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, the share of H-type firms $m_H = 1/3$, the non-rivalry parameter $\iota = 0.6$, the intertemporal discount factor $\beta = 0.96$, the cyberrisk index $\vartheta = 0.75$, the maximum quality threshold $\overline{A} = 25$, the innovation externality $b = 0.035$, and the cost of capital $r = 1$.

The model is simple, but it generates some powerful predictions. data risk hurts firms in the modern economy and firms make lower profits at increasingly high levels of risk. However, there is a silver lining: data risk can sustain growth and innovation when it allows firms to use data security protection for innovation. We allowed some firms in the economy the potential to use data security to improve their productivity ceiling. When given this opportunity, data risk can sustain firm growth and innovation because there are innovation externalities that arise from data risk protection.

5 Conclusion

In this paper, we assess the relationship between data risk, data security, innovation, and growth. From the empirical stand point, we find evidence that the increased threat of data theft and destruction drives innovation in security measures and systems, leading to advancements in technology and potential long-term growth when security measures are developed in-house, in AI-intensive firms. Essentially, the data risk motivates AI-intensive companies to actively pursue digital innovation, subsequently

enhancing productivity in various aspects of their operations.

In other words, AI-intensive firms which develop products and services to protect themselves against data risk, benefit from these products and services to improve the quality of their other digital products. In this context, it is noteworthy to consider how Amazon's innovation with the 1-click purchase system relies on a patented innovation that ensures secure data transmission over the internet. This innovation and its associated patent have not only revolutionized the online shopping experience, but they also highlight the critical role of secure data transmission in the digital realm. Amazon's use of their own internally-developed data-security innovation at the heart of their other digital product offerings aligns with our empirical analysis, confirming that digitally-intensive firms respond to data risk by boosting their innovation activities with positive spillover effects across multiple product domains.

We also find that early treated firms, in the sense of firms being in states that adopted data breach notification laws early, display the strongest response. This suggests that the nature of data risk may have changed over time, becoming more severe and debilitating, increasing firm costs beyond their ability to invest in innovation. But the muted response in the later part of the sample could also be due to the multi-state nature of the Data Breach Notification Laws. Yet, this feature only strengthens our empirical findings, as firms in not-yet-treated states may already react before treatment, implying that the estimates we compute in the difference-in-difference exercise are a lower bound of treated firms' responses. The exact mechanism for this effect is interesting in and of itself and the subject of future investigation.

In this paper we also propose a growth model of the data economy where data, crucial for business optimization, is at risk of being damaged and destroyed. We allow firms to protect themselves against data risk and even trade data security protection. Our simple model features heterogeneity in the type of data security a firm invests in. AI-intensive firms invest in in-house data security, which can be used to improve the quality of other products. Non-AI-intensive firms invest in the external data security they source from AI-intensive firms. This external data protection, which they buy, is assumed to be not tailored enough for them to be used for the development of other products within those firms. Similar to our empirical findings, the model generates growth and innovation for AI-intensive firms, while non-AI intensive firms experience a decrease in profits and innovation activities in response to greater data risk.

In the context of a data-driven economy increasingly threatened by data risk, the divide between high-tech and low-tech firms has become more pronounced, with significant policy implications. AI-intensive firms, benefiting from positive spillovers generated by their innovations in data security, are incentivized to continue their advancements. This dynamic, however, runs the risk of widening the technological and

security gap between them and their low-tech counterparts, potentially exacerbating industry concentration and inequity. Small and Medium Enterprises (SMEs), in particular, find themselves at a disadvantage. Lacking the sophisticated algorithms, data, and specialized personnel available to high-tech firms, SMEs emerge as increasingly vulnerable targets for data security threats. This vulnerability underscores the urgent need for government intervention, both in terms of incentivizing robust data protection measures among SMEs and implementing more effective regulatory frameworks to protect them.

Additionally, there is a critical need to develop a financial insurance market tailored to data security. Currently underdeveloped and costly, such a market could encourage better protection standards across firms through the strategic design of insurance contracts, offering a financial safety net that incentivizes firms to adopt and maintain higher levels of data security. These policy directions not only aim to level the playing field among firms of varying technological capabilities, but also to strengthen the overall resilience of the economic landscape against the growing threat of data risk.

References

Aghion, Philippe, John Van Reenen, and Luigi Zingales, "Innovation and institutional ownership," *American Economic Review*, 2013, *103* (1), 277–304.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Power and Prediction: The Disruptive Economics of Artificial Intelligence*, Harvard Business Press, 11 2022.

Akey, Pat, Stefan Lewellen, and Inessa Liskovich, "Hacking Corporate Reputations," *SSRN Electronic Journal*, 01 2018.

Aldasoro, Iñaki, Leonardo Gambacorta, Paolo Giudici, and Thomas Leach, "The drivers of cyber risk," *Journal of Financial Stability*, Jun 2022, *60*, 100989.

Amore, Mario Daniele, Cedric Schneider, and Alminas Žaldokas, "Credit supply and corporate innovation," *Journal of Financial Economics*, 2013, *109* (3), 835–855.

Aridor, Guy, Yeon-Koo Che, and Tobias Salz, "The effect of privacy regulation on the data industry: empirical evidence from GDPR," *The RAND Journal of Economics*, 2023, *54* (4), 695–730.

Autor, David, David Dorn, Lawrence F Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms*," *Quarterly Journal of Economics*, 02 2020, *135* (2), 645–709.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S Graff Zivin, "Does science advance one funeral at a time?," *American Economic Review*, 2019, *109* (8), 2889–2920.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "Artificial intelligence, firm growth, and product innovation," *Journal of Financial Economics*, 2024, *151*, 103745.

Baker, Andrew C., David F. Larcker, and Charles C.Y. Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Blundell, Richard, Rachel Griffith, and John Van Reenen, "Market share, market value and innovation in a panel of British manufacturing firms," *The Review of Economic Studies*, 1999, *66* (3), 529–554.

Boasiako, Kwabena A. and Michael O'Connor Keefe, "Data breaches and corporate liquidity management," *European Financial Management*, 2021, *27* (3), 528–551.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," Forthcoming in ReStud, 2108.12419, arXiv.org 2022.

Brynjolfsson, Erik, Daniel Rock, and Chad Syverson, "Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics," NBER Working Papers 24001, National Bureau of Economic Research, Inc 2017.

Cohn, Jonathan B, Zack Liu, and Malcolm I Wardlaw, "Count (and count-like) data in finance," *Journal of Financial Economics*, 2022, *146* (2), 529–551.

Correia, Sergio, Paulo Guimarães, and Tom Zylkin, "Fast Poisson estimation with high-dimensional fixed effects," *The Stata Journal*, 2020, *20* (1), 95–115.

Dass, Nishant, Vikram Nanda, and Steven Chong Xiao, "Truncation bias corrections in patent data: Implications for recent research on innovation," *Journal of Corporate Finance*, 2017, *44*, 353–374.

Duffie, Darrell and Jeremy Younger, "Cyber Runs," Hutchins Center Unpublished Working Paper 51, Brookings Institution, Washington, D.C. 2019.

Eeckhout, Jan and Laura Veldkamp, "Data and Market Power," CEPR Discussion Papers 17272, C.E.P.R. Discussion Papers May 2022.

Ewens, Michael, Ryan Peters, and Sean Wang, "Measuring Intangible Capital with Market Prices," *Working Paper*, 2020.

Farboodi, M. and L. Veldkamp, "A Growth Model of the Data Economy," Technical Report 28427 2021.

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

Florackis, Chris, Christodoulos Louca, Roni Michaely, and Michael Weber, "Cybersecurity risk," *The Review of Financial Studies*, 2023, *36* (1), 351–407.

Ganglmair, Bernhard, W Keith Robinson, and Michael Seeligson, "The rise of process claims: Evidence from a century of US patents," *Available at SSRN 4069994*, 2022.

Giczy, Alexander V, Nicholas A Pairolero, and Andrew A Toole, "Identifying artificial intelligence (AI) invention: A novel AI patent dataset," *The Journal of Technology Transfer*, 2022, *47* (2), 476–505.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg, "The NBER patent citation data file: Lessons, insights and methodological tools," 2001.

_ , Adam Jaffe, and Manuel Trajtenberg, "Market value and patent citations," *RAND Journal of economics*, 2005, pp. 16–38.

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches, "Econometric Models for Count Data with an Application to the Patents-R & D Relationship," *Econometrica*, 1984, *52* (4), 909–938.

Hoberg, Gerard and Gordon Phillips, "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, 2016, *124* (5), 1423–1465.

Howell, Sabrina T, "Financing innovation: Evidence from R&D grants," *American Economic Review*, 2017, *107* (4), 1136–64.

Huang, Henry and Chong Wang, "Do Banks Price Firms' Data Breaches?," *The Accounting Review*, 2021, *96* (3), 261–286.

Jamilov, Rustam, Hélène Rey, and Ahmed Tahoun, "The anatomy of cyber risk," Technical Report, National Bureau of Economic Research 2021.

Jones, C.I. and C. Tonetti, "Nonrivalry and the Economics of Data," *American Economic Review*, 2020, *110* (9), 2819–2858.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological innovation, resource allocation, and growth," *The Quarterly Journal of Economics*, 2017, *132* (2), 665–712.

Lerner, Josh and Amit Seru, "The use and misuse of patent data: Issues for finance and beyond," *The Review of Financial Studies*, 2022, *35* (6), 2667–2704.

Liu, Jinyu and Xiaoran Ni, "Ordeal by innocence in the big-data era: Intended data breach disclosure, unintended real activities manipulation," *European Financial Management*, 2023.

Mihet, Roxana and Thomas Philippon, "The economics of big data and artificial intelligence," 2019, pp. 29–43.

Murciano-Goroff, Raviv, "Do Data Breach Disclosure Laws Increase Firms' Investment in Securing Their Digital Infrastructure?," in "in."

Perkins Coie LLP, "Security Breach Notification Chart," https://www.perkinscoie.com/en/news-insights/security-breach-notification-chart.html 2023. Accessed on: 2023-06-29.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer, "Regulatory Spillovers and Data Governance: Evidence from the GDPR," *Marketing Science*, jul 2022, *41* (4), 746–768.

Silva, JMC Santos and Silvana Tenreyro, "The log of gravity," *The Review of Economics and statistics*, 2006, *88* (4), 641–658.

_ and _ , "Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator," *Economics Letters*, 2011, *112* (2), 220–222.

Wooldridge, Jeffrey M, "Distribution-free estimation of some nonlinear panel data models," *Journal of Econometrics*, 1999, *90* (1), 77–97.

[12pt,letter]article

[left=3cm, right=3cm, top=3cm, bottom=3cm]geometry

setspace,amsfonts,comment,amsmath,amssymb,amsxtra,ushort,tikz,graphicx,pgfplots,units,xfrac hyperref lscape afterpage natbib setspace graphicx subcaption [caption]subfig setspace amsmath kantlipsum [title]appendix longtable longtable,threeparttablex setspace comment float

[normalem]ulem

epigraph

etoolbox

csquotes

hyperref comment [final]pdfpages tikz amssymb bbold longtable tabularx

pdflscape threeparttable booktabs,caption rotating multirow

hyperref

Online Appendix

Appendix A Data risk scores

A.1 Florakis et al. (2023) method

Our algorithm extracts cybersecurity risk discussions from the "Item 1A. Risk Factors" section of 10-K filings available on the SEC Edgar database, excluding amendments. It uses a web-crawling algorithm to download and examine filings for the fiscal year and central index key (CIK) details, alongside cybersecurity risk disclosures. To capture direct cybersecurity risk discussions, a list of specific keywords (e.g., "unauthorized access", "attack", "hacker") is compiled. The presence or absence of related or unrelated keywords within the same sentence refines the extraction process, aiming to filter out noise and ensure relevancy.

Indirect descriptions of cybersecurity risk, such as details about a firm's business, security measures, and the potential consequences of a cyberattack, are also considered. These descriptions are categorized into internal, legal, and economic consequences. Another list of indirect keywords/phrases helps retrieve sentences relevant to cybersecurity risk, requiring an initial sentence with direct risk discussion before searching for indirect cues in the subsequent 10 sentences.

Developing a firm-level measure of cybersecurity risk poses challenges, particularly because disclosures often blend discussions of business operations with cybersecurity risk exposure and management strategies. To create a more accurate measure, the study isolates firms subject to major cyberattacks as a training sample. This approach aims to capture a firm's exposure to cybersecurity risk by comparing its disclosures to those of the training sample, focusing on systematic components of risk.

The training sample comprises firms that have experienced "major" cyberattacks, defined by their coverage in global news outlets and the impact on lost personal information through hacking or malware. A total of 175 cyberattacks identified between 2005 and 2018, with

69 classified as "major," inform the training sample. The construction of the cybersecurity risk measure involves comparing the textual information in a firm's disclosures to those in the training sample, excluding certain types of words to focus on word roots. Similarity measures, both cosine and Jaccard, assess the overlap between textual vectors of disclosures, aiming to quantify cybersecurity risk exposure.

We use the huge list of 3210 words in the dictionary Florakis et al. (2023) provided us. Notice Florakis et al. (2023) select the following 10 sentences after a relevant word is hit. But they stop if they encounter a BOLD font (for a new risk-factor other than cyber-security). We do not have BOLD font, so we need to find other rule to stop. Instead of using "BOLD" to stop the algo after 10 sentences, we can use "Risks relating to" or "Risks specific to" or "risks related to" or "Risk relating to" or "Risk specific to" or "risk related to" in a new individual line.

Below we elaborate on the keyword/phrases used to extract relevant discussions on cybersecurity risk, distinguishing between direct and indirect descriptions. The approach considers language nuances and the context in which keywords appear, refining the search to relevant discussions only.

I. Direct description	**Relevant hit if**	**Irrelevant hit if**
Threat	Cyber-, cyber, networks, systems, products, services, datacenter, infrastructure	Terror, war, contraband, bombs
Attack	Cyber-, cyber, networks, systems, products, services, datacenter, infrastructure	Terror, simulator, disease, legal action, competitive, competitors, substitute, patent, nuclear, life, threaten/ed
Computer, information system	Malware, virus, viruses, intrusions Software, programs, third parties, attacks	fires, product sales, warranty claim/s Fiduciary duty/duties, covenant/s, credit, agreement/s, warranty, warranties, obligations, regulations, contract/s, resolution
Breaches		
Malicious		
Hacker, hacking, social engineering, denial of service, denial-of-service, phishing, cyberattack, cyberattacks, cyber risk, cyber security, cybersecurity, cyber intrusions, unauthorized access, breach in security, security breach		
Indirect description		
2.1 Company Business		
Company, regular course	Business, operation, services	
Technology, technologies	Computer, information, communication, proprietary, infrastructure, reliance, digital, advances	

Information	Network, services, systems, confidential, proprietary, account	
Electronic	Network, services, systems, information	
Computer, telecommunication, third-party, infrastructure	Systems, networks, facilities	
Collect, store, transmit, retrieve, sensitive, critical, protection	Data, information	
IT environment, IT systems, operational systems, communication systems, critical infrastructure		
Security	Network, products, services, systems, devices, data, infrastructure, patches, cloud, web, email, vulnerabilities, threat, breach, penetrate, bypass, compromised, incidence, incident, circumvent, measures, portfolio, solutions, practices, standards	
Vulnerabilities	Network, products, services, systems, devices, data, infrastructure, claims	
2.2. Internal consequences		
Integrity, reliability, protect, protection, protecting, prevent, prevention, preventing, monitors, compromise, secure, failure	Network, products, services, systems, devices, data, measures, information	
Gain access	Network, systems, data, datacenter	

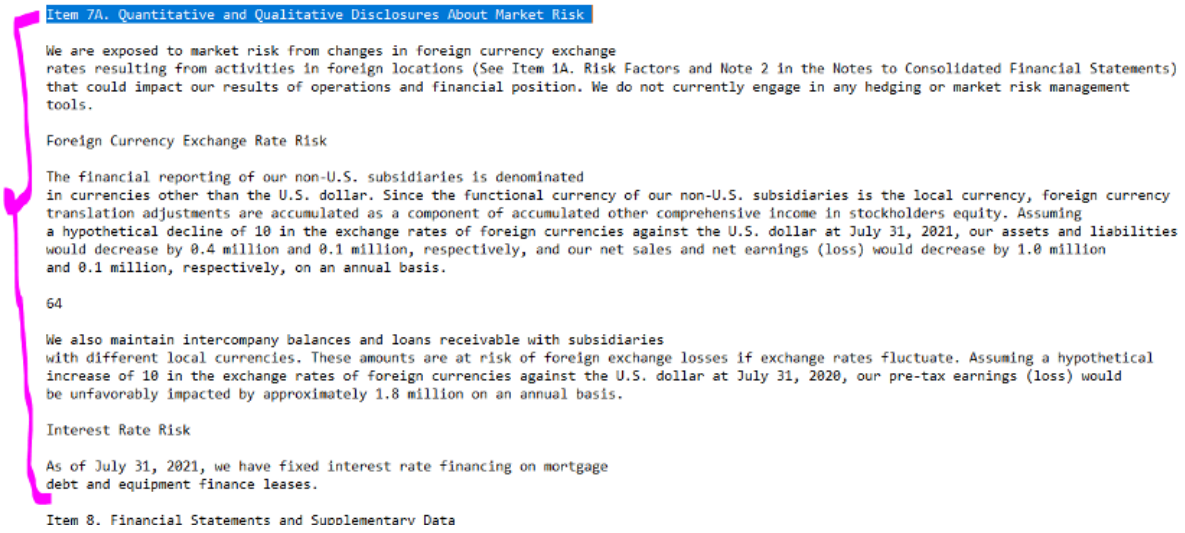
Access, accessed, modified	Improper, improperly	
Theft, misuse, misusing, modification, destruction, lost, loss, stolen, steal, disclose, publicly disclosed	Assets, intellectual property, data, information	
Investigate, remediate, remediation, recover, repair, replace	Network, products, services, systems, data, measures, efforts	
Interruptions, disruptions, delays	Network, services, system	
Degrade the user experience, invasion, user names, password, break-ins, terminated agreements		
2.3. Legal consequences		
Legal	Claims, actions, challenges, liability	
Legislative	Actions	
Regulatory	Actions, investigations, agencies	
Liability	Claims	
Lawsuits, litigation		
2.4. Economic consequences		
Business	Adversely, material, harm disruptive, negative	
Operations, services	Disrupt	
Revenues	Reduce, adversely, loss, lose	
Cost	Increase, increasing, remedy	
Operating results, operating margin	Harm, diminish, reduce	
Earnings	Reduce, adversely	

Financial	Harm, diminish, adversely, material, damage, negative	
Competitive position	Harm, diminish	
Reputation	Harm, damage, loss, adverse	
Brand	Harm, damage	

A.2 Robustness

We check the robustness of our method above by creating our own scores with our own dictionary.

We extract the entire Item 1A. Risk Factors and Item 7 (we do not use the title of Item 7 because it differs between reports; also sometimes it is written as Item 7A, or Item 7, just use the entire section Item 7 until the start of Item 8.).



Item 7A. Quantitative and Qualitative Disclosures About Market Risk

We are exposed to market risk from changes in foreign currency exchange rates resulting from activities in foreign locations (See Item 1A. Risk Factors and Note 2 in the Notes to Consolidated Financial Statements) that could impact our results of operations and financial position. We do not currently engage in any hedging or market risk management tools.

Foreign Currency Exchange Rate Risk

The financial reporting of our non-U.S. subsidiaries is denominated in currencies other than the U.S. dollar. Since the functional currency of our non-U.S. subsidiaries is the local currency, foreign currency translation adjustments are accumulated as a component of accumulated other comprehensive income in stockholders equity. Assuming a hypothetical decline of 10 in the exchange rates of foreign currencies against the U.S. dollar at July 31, 2021, our assets and liabilities would decrease by 0.4 million and 0.1 million, respectively, and our net sales and net earnings (loss) would decrease by 1.0 million and 0.1 million, respectively, on an annual basis.

64

We also maintain intercompany balances and loans receivable with subsidiaries with different local currencies. These amounts are at risk of foreign exchange losses if exchange rates fluctuate. Assuming a hypothetical increase of 10 in the exchange rates of foreign currencies against the U.S. dollar at July 31, 2020, our pre-tax earnings (loss) would be unfavorably impacted by approximately 1.8 million on an annual basis.

Interest Rate Risk

As of July 31, 2021, we have fixed interest rate financing on mortgage debt and equipment finance leases.

Item 8. Financial Statements and Supplementary Data

Now, instead of populating the huge list of 3210 words (given in the file WordRoots.xlsx), only populate the following short list of 160 cybercrime specific words:

Terms	**Source**
Access Control, Attacker, Authentication, Cloud Safety, Cloud Security, Computer Breach, Computer Security, Computer Virus, Cyber Attack, Cyber Incident, Cyber Security, Cyber Threat, Cybersecurity, Data Breach, Data Integrity, Data Leakage, Data Loss, Data Security, Data Theft, Digital Security, Encrypt, Encryption, Exfiltration, Exposed Data, Firewall, Hacker, Information Leak, Information Risk, Information Security, Information System, Infosec, Inside Threat, Insider Threat, Intrusion, IT Asset, Malicious, Malware, Network Resilience, Phishing, Security Breach, Security Expenditure, Security Incident, Security Integrity, Security Measure, Security Monitoring, Security Policy, Security Program, Software Assurance, Spoof, Spyware, System Integrity, System Security, Unauthorized Access	Ma(2023)paper

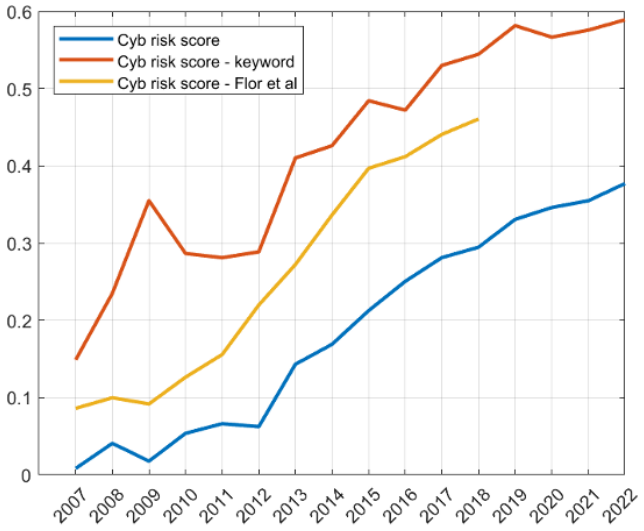
breach in security, cyber intrusions, cyber risk, cyber-attack, cyberattacks, denial of service, denial-of-service, hacking, social engineering	Weber(2022)paper
antispyware, antivirus, blocklist, Confidentiality, Cookie, decipher, decode, decrypt, Decryption, encode, Hash, hashing, password, protocol, Proxy, Public-Key, Ransomware, Spam, information theft, Trojan, Two-factor, Vulnerability Assessment, Vulnerability Management, Vulnerability Scanning, Vulnerability Exploit, Vulnerability Patching, Vulnerability Analysis, Software Vulnerability, Network Vulnerability, System Vulnerability, Security Vulnerability, Vulnerability Disclosure, Zero-Day Vulnerability, Remote Code Execution Vulnerability, Database Vulnerability, Application Vulnerability, Penetration Testing, Security Auditing, Intrusion Detection System, Intrusion Prevention System, Secure Sockets Layer, Transport Layer Security, Virtual Private Network, Firewall Configuration, Multi-Factor Authentication, Brute Force Attack, Security Architecture, Endpoint Protection, Security Operations Center, Security Information and Event Management, Risk Assessment, Data Encryption Standard, Advanced Encryption Standard, Public Key Infrastructure, Domain Name System Security, Secure/Multipurpose Internet Mail Extensions, Botnet, Social Engineering Attack, Spear Phishing, Zero Trust Architecture, Security Compliance, Secure Coding, Sandboxing, Threat Intelligence, Incident Response Plan, Chain of Custody, Data Masking, Digital Forensics, Man-in-the-Middle Attack, Dark Web Monitoring, Certificate Authority, Secure File Transfer Protocol (SFTP), Endpoint Detection and Response (EDR), Web Application Firewall (WAF), Privilege Escalation	BSIGroup/NICCS/SANS

computer attacks, computer intrusion, computer malware, cyber threats, Data stealing, datacenter attack, digital breach, digital leak, digital loss, information system attacks, infrastructure attack, Network attack, Network integrity, Network security, network threats, programs breach, services threat, software breach, System attack, system threat, Systems attack, systems threat, third party breach	Own

A.3 Correlation between methods with original Florakis et al. (2023) score

The correlation between these methods can be seen in the figure below:

Figure 20: Comparison between our extended scores and Florakis et al. (2023) score



The correlation between our extended score (in blue) with the Florakis et al. (2023) cyber-risk score (in yellow) is 75%. The difference between our methods is using another list of publicly breached firms, but keepiong the same large dictionary of words as explained above. The correlation between our extended score based on a smaller dictionary of cyber security risk (in red) and the Florakis et al. (2023) cyber-risk score is 64%.

Appendix B Robustness

This section provides robustness checks for tables and figures included in the paper. We run these regressions using different controls.

B.1 Robustness checks for reduced-form analysis

Table 8: Regression of financial variables

	Log assets	ROA	Tobin's q	Book-to-market	Leverage	Tangibility	COGS-to-asset	Opex-to-asset
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lagged cyberrisk score*(AI = 0)	0.0798	0.0189	-0.341**	-0.0507	0.00506	-0.00834	-2.214	-0.135***
	(0.0509)	(0.0174)	(0.159)	(0.0454)	(0.0160)	(0.00697)	(3.610)	(0.0290)
Lagged cyberrisk score*(AI = 1)	0.249***	0.0811***	-0.494**	0.0321	-0.00151	0.00171	-12.66*	0.0264
	(0.0709)	(0.0276)	(0.206)	(0.0556)	(0.0246)	(0.00781)	(6.558)	(0.0377)
Lagged controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	20238	20234	20238	20238	20237	20238	19409	17990

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. The coefficient estimates are derived from the OLS estimate. Standard errors are denoted in parentheses and are clustered at the firm level. Here, N represents the total number of firm-year observations. Cyber score, and control variables are lagged by one year. Cyberrisk score is constructed using the methodology Florackis et al. (2023). Each regression controls for financial variables, except for the variable which is the dependent variable in the regression. These control variables include: Log of total assets, log of R&D expenditure, Tobin's Q, Return On Assets, Tangibility, Leverage, Book-to-market ratio, Cash-to-asset ratio. For variable description see table **??**.

B.2 Robusness checks for DiD analysis

Table 9: Regression of citation-weighted patent count by industry

	Citation-weighted patent count	
	(1)	(2)
Lagged cyberrisk score*(Ind = Agri)	-0.473***	-0.552***
	(0.136)	(0.147)
Lagged cyberrisk score*(Ind = Mining)	-0.657	-0.436
	(0.484)	(0.489)
Lagged cyberrisk score*(Ind = Manu-I)	1.076	1.920***
	(0.757)	(0.609)
Lagged cyberrisk score*(Ind = Manu-II)	-0.0141	0.0720
	(0.230)	(0.236)
Lagged cyberrisk score*(Ind = Manu-III)	0.294*	0.321**
	(0.159)	(0.159)
Lagged cyberrisk score*(Ind = Wholesale)	-0.508	1.778***
	(0.823)	(0.440)
Lagged cyberrisk score*(Ind = Retail)	-0.629	0.257
	(0.671)	(1.069)
Lagged cyberrisk score*(Ind = Information)	0.00411	0.0585
	(0.325)	(0.354)
Lagged cyberrisk score*(Ind = Fin and Ins)	2.595***	-1.839***
	(0.643)	(0.231)
Lagged cyberrisk score*(Ind = Real estate)	0.363	0.314
	(0.505)	(0.401)
Lagged cyberrisk score*(Ind = Prof Svs)	1.747*	1.559
	(0.971)	(0.968)
Lagged cyberrisk score*(Ind = Admin)	1.196	0.121
	(0.896)	(1.971)
Lagged cyberrisk score*(Ind = Healthcare)	0.0706	-1.848
	(0.609)	(2.031)
Lagged cyberrisk score*(Ind = Accomodation)	-0.846**	-0.468*
	(0.367)	(0.266)
Lagged cyberrisk score*(Ind = Unclassified)	0.0552	-0.0214
	(0.0787)	(0.0833)
Lagged Log (R&D Expenditure)		0.281***
		(0.0920)
Lagged controls	Yes	Yes
Firm FE	Yes	Yes
Year FE	Yes	Yes
N	16539	13343

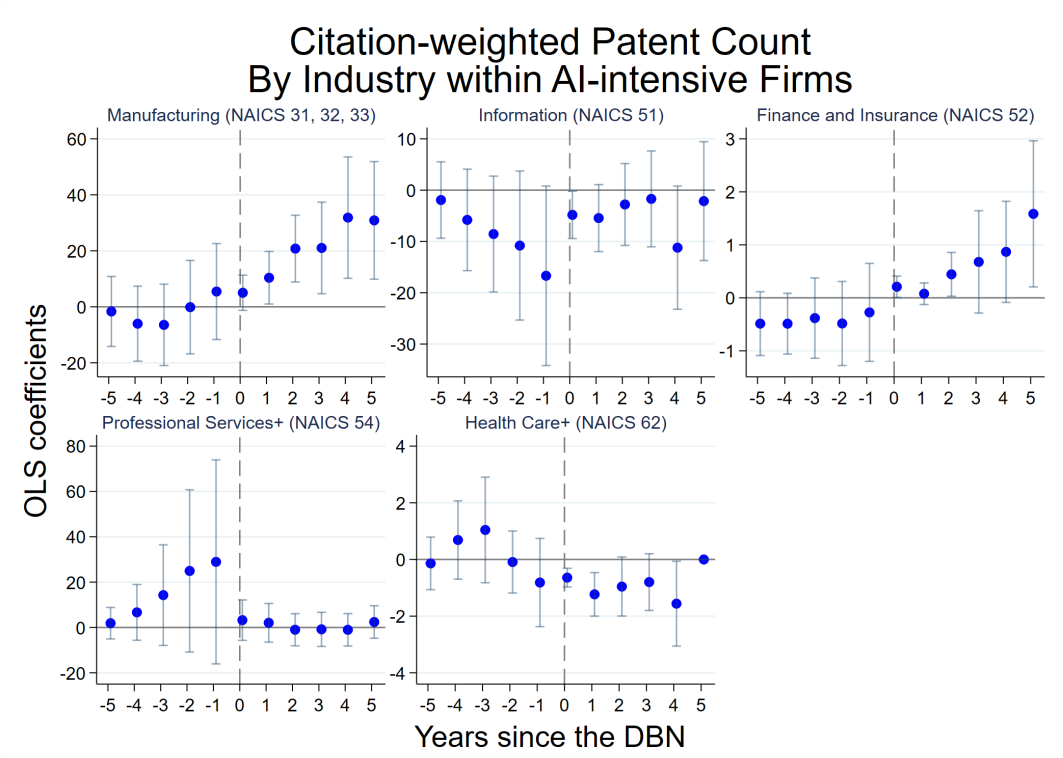
$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. The coefficient estimates are derived from the Poisson pseudo-maximum likelihood estimation. Standard errors are denoted in parentheses and are clustered at the firm level. Here, N represents the total number of firm-year observations. Cyber score, and control variables are lagged by one year. Cyberrisk score is constructed using the methodology Florackis et al. (2023). Citation-weighted patent count weighs each patent with the forward citation the patent receives, adjusting for the filing vintage. Industry refers to North American Industry Classification System (NAICS) at the two-digit level. Following 2-digit industry codes were part of analysis: 11 (Agriculture, Forestry, Fishing and Hunting), 21 (Mining, Quarrying, and Oil and Gas Extraction), 31 (Manufacturing-I), 32 (Manufacturing-II), 33 (Manufacturing-III), 42 (Wholesale trade), 45 (Retail Trade), 51 (Information), 52 (Finance and Insurance), 53 (Real Estate and Rental and Leasing), 54 (Professional, Scientific, and Technical Services), 56 (Admin, Support, Waste Management, Remediation Services), 62 (Health Care and Social Assistance), 72 (Accommodation and Food Services), 99 (Unclassified). Control variables include: Log of total assets, log of R&D expenditure, Tobin's Q, Return On Assets, Tangibility, Leverage, Book-to-market ratio, Cash-to-asset ratio. For variable description see table **??**.

Table 10: Regression of citation-weighted patent count within superstar firms

	2007 Vintage		2000 Vintage	
	Top 100	Top 500	Top 100	Top 500
	(1)	(2)	(3)	(4)
Lagged cyberrisk score*(AI = 0)	-0.0351	0.221	0.197	-0.0302
	(0.166)	(0.216)	(0.177)	(0.148)
Lagged cyberrisk score*(AI = 1)	0.558**	0.311	0.399*	0.409**
	(0.231)	(0.222)	(0.214)	(0.197)
Lagged controls	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	552	1874	368	1470

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. The coefficient estimates are derived from the Poisson pseudo-maximum likelihood estimation. Standard errors are denoted in parentheses and are clustered at the firm level. Here, N represents the total number of firm-year observations. Cyber score, and control variables are lagged by one year. Cyberrisk score is constructed using the methodology Florackis et al. (2023). The dependent variable, Citation-weighted patent count weighs each patent with the forward citation the patent receives, adjusting for the filing vintage. Analysis conducted on the highest N firms in terms of total sales for a given vintage year, where N equals either 100 or 500. These firms are referred to as 'superstar firms,' following the terminology used by Autor et al. (2020). Control variables include: Log of total assets, log of R&D expenditure, Tobin's Q, Return On Assets, Tangibility, Leverage, Book-to-market ratio, Cash-to-asset ratio. For variable description see table 1.

Figure 21: Citation-weighted patent count within AI-intensive firms: by select industries



This figure plots BJS dynamic heterogeneous treatment effects of citation-weighted patent counts. Industry refers to North American Industry Classification System (NAICS) at the two-digit level. The '0' event is the staggered adoption of DBN laws across the United States. Estimates for AI-intensive firms. AI intensive firms are identified using a combination of the USPTO dataset on AI patents (Giczy et al. (2022)) and the KPSS patent dataset linked to firms (Kogan et al. (2017)). Moreover, firms that are close to AI patenting firms in the sense of Hoberg and Phillips (2016) and mirror their AI innovations are also considered AI-intensive. This measure has the advantage to be constructed from entirely publicly available data and it is different from IT expenditures. The figure plots the estimates only for those industries where a separate estimation was possible for the AI intensive firms.

Appendix C Theoretical derivations

C.1 Model Solution Details

There are two sources of uncertainty in firm i's problem at date t: the (random) optimal technique $\theta_{i,t}$, and the aggregate price P_t. Let $(\hat{\mu}_{i,t}, \Omega_{i,t})$ denote the conditional mean and precision of firm i belief about $\theta_{i,t}$ given its information set at date t, $\mathcal{I}_{i,t}$.

In this section, we will first describe the firm belief updating process about its optimal technique. Next, we argue that in this environment, the firm's optimal production choice is deterministic, and thus the price is deterministic as well. Finally, we lay out the full set of equations that characterize the equilibrium of this economy with two groups of firms.

Belief updating The information problem of firm i about its optimal technique $\theta_{i,t}$ can be expressed as a Kalman filtering system, with a 2-by-1 observation equation, $(\hat{\mu}_{i,t}, \Omega_{i,t})$.

We start by describing the Kalman system, and show that the sequence of conditional variances is deterministic. Note that all the variables are firm specific, but since the information problem is solved firm-by-firm, for brevity we suppress the dependence on firm index i.

At time t, each firm observes two types of signals. First, date $t-1$ output provides a noisy signal about θ_{t-1}:

$$y_{t-1} = \theta_{t-1} + \epsilon_{a,t-1}, \tag{26}$$

where $\epsilon_{a,t} \sim \mathcal{N}(0, \sigma_a^2)$. We provide model detail on this step below. Second, the firm observes $n_t = z_t$ data points as a bi-product of its economic activity. The set of signals $\{s_{t,m}\}_{m\in[1:n_{i,t}]}$ are equivalent to an aggregate (average) signal $\bar{s}_t$ such that:

$$\bar{s}_t = \theta_t + \epsilon_{s,t}, \tag{27}$$

where $\epsilon_{s,t} \sim \mathcal{N}(0, \sigma_\epsilon^2/n_t)$. The state equation is

$$\theta_t - \bar{\theta} = \rho(\theta_{t-1} - \bar{\theta}) + \eta_t,$$

where $\eta_t \sim \mathcal{N}(0, \sigma_\theta^2)$.

At time, t, the firm takes as given:

$$\begin{aligned}\widehat{\mu}_{t-1} &= \mathbb{E}\big[\theta_t \mid s^{t-1}, y^{t-2}\big] \\ \Omega_{t-1}^{-1} &= Var\big[\theta_t \mid s^{t-1}, y^{t-2}\big]\end{aligned}$$

where $s^{t-1} = \{s_{t-1}, s_{t-2}, \dots\}$ and $y^{t-2} = \{y_{t-2}, y_{it-3}, \dots\}$ denote the histories of the observed variables, and $s_t = \{s_{t,m}\}_{m\in[1:n_{i,t}]}$.

We update the state variable sequentially, using the two signals. First, combine the

priors with y_{t-1}:

$$\begin{aligned}
\mathbb{E}\left[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\right] &= \frac{\Omega_{t-1}\widehat{\mu}_{t-1} + \sigma_a^{-2} y_{t-1}}{\Omega_{t-1} + \sigma_a^{-2}} \\
V\left[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\right] &= \left[\Omega_{t-1} + \sigma_a^{-2}\right]^{-1} \\
\mathbb{E}\left[\theta_t \mid \mathcal{I}_{t-1}, y_{t-1}\right] &= \bar{\theta} + \rho \cdot \left(\mathbb{E}\left[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\right] - \bar{\theta}\right) \\
V\left[\theta_t \mid \mathcal{I}_{t-1}, y_{t-1}\right] &= \rho^2 \left[\Omega_{t-1} + \sigma_a^{-2}\right]^{-1} + \sigma_\theta^2
\end{aligned}$$

Then, use these as priors and update them with $\bar{s}_t$:

$$\widehat{\mu}_t = \mathbb{E}\left[\theta_t \mid \mathcal{I}_t\right] = \frac{\left[\rho^2 \left[\Omega_{t-1} + \sigma_a^{-2}\right]^{-1} + \sigma_\theta^2\right]^{-1} \cdot \mathbb{E}\left[\theta_t \mid \mathcal{I}_{t-1}, y_{t-1}\right] + n_t \sigma_\epsilon^{-2} \bar{s}_t}{\left[\rho^2 \left[\Omega_{t-1} + \sigma_a^{-2}\right]^{-1} + \sigma_\theta^2\right]^{-1} + n_t \sigma_\epsilon^{-2}} \tag{28}$$

$$\Omega_t^{-1} = Var\left[\theta \mid \mathcal{I}_t\right] = \left\{\left[\rho^2 \left[\Omega_{t-1} + \sigma_a^{-2}\right]^{-1} + \sigma_\theta^2\right]^{-1} + n_t \sigma_\epsilon^{-2}\right\}^{-1} \tag{29}$$

Multiply and divide equation (28) by Ω_t^{-1} as defined in equation (29) to get

$$\widehat{\mu}_t = (1 - n_t \sigma_\epsilon^{-2} \Omega_t^{-1}) \left[\bar{\theta}(1-\rho) + \rho\left((1 - M_t)\mu_{t-1} + M_t \tilde{y}_{t-1}\right)\right] + n_t \sigma_\epsilon^{-2} \Omega_t^{-1} \bar{s}_t, \tag{30}$$

where $M_t = \sigma_a^{-2}(\Sigma_{t-1} + \sigma_a^{-2})^{-1}$.

Equations (29) and (30) constitute the Kalman filter describing the firm dynamic information problem. Importantly, note that Ω_t^{-1} is deterministic.

Appendix D Modeling quadratic-normal signals from output

When y_{t-1} is observed, agents can back out A_{t-1} exactly. To keep the model simple, we assumed that when agents see A_{t-1}, they also learn whether the quadratic term $(a_{t-1} - \theta_{t-1} - \epsilon_{a,t-1})^2$ had a positive or negative root. An interpretation is that they can figure out if their action a_t was too high or too low.

Relaxing this assumption complicates the model because, when agents do not know which root of the square was realized, the signal is no longer normal. One might solve a model with binomial distribution over two normal variables, perhaps with other simplifying assumptions. For numerical work, a good approximate solution would be to simulate the binomial-normal and then allows firms to observe a normal signal with the same mean and same variance as the true binomial-normal signal. This would capture the right amount of information flow, and keep the tractability of updating with normal variables.

Appendix E The cybersecurity planning problems: optimality conditions and steady state results

E.1 H-type firm

The current-value Hamiltonian function for the H-type firm:

$$H(\Omega_{H,t};\tau_t;\delta_t;p_{H,t}) = \Pi_{H,t,\text{sec}}$$
$$+\beta p_{H,t+1}\left\{\left[\rho^2(\Omega_{H,t}+\sigma^{-2})^{-1}+\sigma^2\right]^{-1}+\left[1-\vartheta e^{-\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}\right]z\sigma^{-2}-\Omega_{i,t}\right\} \tag{31}$$

where $p_{H,t}$ is the shadow-price or co-state variable associated with the state variable. The transversality condition is $\lim_{t\to\infty}\Omega_{H,t}\beta^t p_{H,t}=0$.

The first-order optimality conditions:

$$\frac{\partial H}{\partial \tau_t}=0\Rightarrow \beta p_{H,t+1}=\frac{1-b\overline{A}e^{b\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}}{\vartheta e^{-\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}z\sigma^{-2}} \tag{32}$$

$$\frac{\partial H}{\partial \delta_t}=0\Rightarrow \beta p_{H,t+1}=\frac{\pi_t-b\overline{A}\iota e^{b\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}}{\vartheta\iota e^{-\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}z\sigma^{-2}} \tag{33}$$

$$\beta p_{H,t+1}-p_{H,t}=-\frac{\partial H}{\partial \Omega_{H,t}}\Rightarrow\left[\rho+\frac{\sigma^2}{\rho}(\Omega_{H,t}+\sigma^{-2})\right]^{-2}\beta p_{H,t+1}=p_{H,t}-\Omega_{H,t}^{-2} \tag{34}$$

From (32) and (33), it emerges a constant optimal trading price, which is simply $\pi_t=\iota$. The price of protection is directly associated with the degree of its own nonrivalry. If protection is completely non-rival (i.e., $\iota=0$), then its price is zero; if protection is fully rival, its price is 1.

Replacing (32) into (34), and evaluating in the steady state, one gets:

$$\Gamma_H=\frac{\vartheta z e^{-\left(\tau^*-\iota\frac{1-u}{u}\delta^*\right)}}{1-b\overline{A}e^{b\left(\tau^*-\iota\frac{1-u}{u}\delta^*\right)}}, \tag{35}$$

with Γ_H defined as $\Gamma_H\equiv\left\{\frac{1}{\beta}-\left[\rho+\frac{\sigma^2}{\rho}(\Omega_H^*+\sigma^{-2})\right]^{-2}\right\}(\Omega_H^*)^2\sigma^2$.

Given constraint (16), it is also true, for the H-firms:

$$\Xi_H=\left[1-\vartheta e^{-\left(\tau^*-\iota\frac{1-u}{u}\delta^*\right)}\right]z, \tag{36}$$

with $\Xi_H\equiv\left\{\Omega_H^*-\left[\rho^2(\Omega_H^*+\sigma^{-2})^{-1}+\sigma^2\right]^{-1}\right\}\sigma^2$.

Combining expressions (35) and (36), one obtains a steady state relation that allows for

the derivation of Ω_H^*:

$$\Gamma_H = \frac{z - \Xi_H}{1 - b\overline{A}\left(\frac{\vartheta z}{z-\Xi_H}\right)^b} \tag{37}$$

Γ_H is such that if $\Omega_H^* = 0$ then $\Gamma_H = 0$ and if $\Omega_H^* \to +\infty$ then $\Gamma_H \to +\infty$.

Ξ_H is such that if $\Omega_H^* = 0$ then $\Xi_H = -\frac{1}{1+\rho^2}$ and if $\Omega_H^* \to +\infty$ then $\Xi_H \to +\infty$. Hence, if $\Omega_H^* = 0$ then $\frac{z-\Xi_H}{1-b\overline{A}\left(\frac{\vartheta z}{z-\Xi_H}\right)^b} = \frac{z+\frac{1}{1+\rho^2}}{1-b\overline{A}\left(\frac{\vartheta z}{z+\frac{1}{1+\rho^2}}\right)^b}$; this is a positive value for $b\overline{A}\left(\frac{\vartheta z}{z+\frac{1}{1+\rho^2}}\right)^b < 1$. If $\Omega_H^* \to +\infty$ then $\frac{z-\Xi_H}{1-b\overline{A}\left(\frac{\vartheta z}{z-\Xi_H}\right)^b} \to -\infty$.

By combining the above reasoning, as long as $b\overline{A}\left(\frac{\vartheta z}{z+\frac{1}{1+\rho^2}}\right)^b < 1$, the l.h.s. of (37) (positively sloped) will intersect the r.h.s. of (37) (negatively sloped) at one single point, and therefore a unique Ω_H^* is derived.

Thus, condition $b\overline{A}\left(\frac{\vartheta z}{z+\frac{1}{1+\rho^2}}\right)^b < 1$ must hold, which can be rewritten as a constraint on ϑ: $\vartheta < \frac{z+\frac{1}{1+\rho^2}}{z}\left(b\overline{A}\right)^{-1/b}$. Because $\vartheta \leq 1$, this constraint is always satisfied as long as $b\overline{A} < 1$.

From (36) also note that the value of security that firm H holds after trade is also a unique constant value,

$$\tau^* - \iota\frac{1-u}{u}\delta^* = -\ln\left(\frac{z-\Xi_H}{\vartheta z}\right) \tag{38}$$

E.2 L-type firm

Turning to the L-type firm, the current-value Hamiltonian is:

$$H(\Omega_{L,t}; \delta_t; p_{L,t}) = \Pi_{L,t,\text{sec}}$$
$$+\beta p_{L,t+1}\left\{\left[\rho^2(\Omega_{L,t}+\sigma^{-2})^{-1}+\sigma^2\right]^{-1} + \left(1-\vartheta e^{-\delta_t}\right)z\sigma^{-2} - \Omega_{i,t}\right\} \tag{39}$$

The transversality condition: $\lim_{t\to\infty}\Omega_{L,t}\beta^t p_{L,t} = 0$.

The first-order conditions are:

$$\frac{\partial H}{\partial \delta_t} = 0 \Rightarrow \beta p_{L,t+1} = \frac{\pi_t}{\vartheta e^{-\delta_t} z\sigma^{-2}} \tag{40}$$

$$\beta p_{L,t+1} - p_{L,t} = -\frac{\partial H}{\partial \Omega_{L,t}} \Rightarrow \left[\rho + \frac{\sigma^2}{\rho}(\Omega_{L,t}+\sigma^{-2})\right]^{-2}\beta p_{L,t+1} = p_{L,t} - \Omega_{L,t}^{-2} \tag{41}$$

Replace (40) into (41), and recall that we already know that $\pi_t = \iota$. With this information, the following steady state condition holds:

$$\Gamma_L = \vartheta z e^{-\delta^*}, \tag{42}$$

with $\Gamma_L \equiv \left\{\frac{1}{\beta} - \left[\rho + \frac{\sigma^2}{\rho}(\Omega_L^* + \sigma^{-2})\right]^{-2}\right\}(\Omega_L^*)^2\sigma^2$.

Given constraint (18),

$$\Xi_L = \left(1 - \vartheta e^{-\delta^*}\right) z, \tag{43}$$

with $\Xi_L \equiv \left\{\Omega_L^* - \left[\rho^2(\Omega_L^* + \sigma^{-2})^{-1} + \sigma^2\right]^{-1}\right\}\sigma^2$.

From (42) and (43), a simple expression emerges for the determination of Ω_L^*,

$$\Gamma_L = z - \Xi_L \tag{44}$$

Equation (44) allows for the derivation of a unique Ω_L^*, because the l.h.s. of the expression is a continuous increasing function starting at zero and diverging to infinity (as Ω_L increases) and the l.h.s. is a continuous decreasing function starting at a positive value and falling to minus infinity (as Ω_L increases).

From (43), one can also compute the steady state value of the amount of security bought by firm L:

$$\delta^* = -\ln\left(\frac{z - \Xi_L}{\vartheta z}\right) \tag{45}$$

A unique δ^* exists as well.

By now, we have computed all the relevant steady state values: Ω_H^* and Ω_L^*, and also δ^* (determined from the L-firm problem), and τ^*, determined from (38) after knowing δ^* (the H-type only decides how much to invest in cyberprotection after knowing how much protection firms in the L sector are willing to buy at price $\pi_t = \iota$).

E.3 Steady-state

Possible steady state scenarios:

- (*i*) The cybersecurity optimal result is such that $\tau^* \leq 0$: firms H do not invest in cybersecurity $\tau^* = 0$ and firms L have no cyberprotection to buy, $\delta^* = 0$. Firms face the problem with no security and their profits are: $\Pi_{H,cy}^* = \Pi_{L,cy}^*$.

- (*ii*) The cybersecurity optimal result is such that $\tau^* > 0$, $\delta^* \leq 0$: firms L will not buy any protection and face the no-protection problem, with profits $\Pi_{L,cy}^*$. Firms of the H type have two possibilities: to invest τ^*, even though they cannot optimally exchange protection, or not to invest; they compare profits $\Pi_{H,\text{sec}}^*$ and $\Pi_{H,cy}^*$ and choose the option that delivers the highest profits.

- (*iii*) The cybersecurity optimal result is such that $\tau^* > 0$ and $\delta^* > 0$: firms find it optimal to invest a positive value in cybersecurity (H) and to trade a positive amount of cybersecurity. In this case, the best option is the cybesecurity one with profits $\Pi_{H,\text{sec}}^*$ and $\Pi_{L,\text{sec}}^*$

Note that conditions $\tau^* > 0$ and $\delta^* > 0$ impose relevant constraints on parameter values, namely, in the first case, $z > \Xi_H$ and $\vartheta > \frac{z - \Xi_H}{z}$ and, in the second case, $z > \Xi_L$ and

$\vartheta > \frac{z-\Xi_L}{z}$. These results suggest that investment and trading in cybersecurity require the cybercrime index ϑ to be above a given threshold.

E.4 Comparative statics

A few intuitive comparative statics outcomes (in the cybersecurity setting, i.e., for $\tau^* > 0$, $\delta^* > 0$):

- (i) $\Delta z > 0$: l.h.s. of (37) does not shift; r.h.s. of (37) shifts right $\Rightarrow$ higher Ω_H^* / l.h.s. of (44) does not move; r.h.s. of (44) shifts right $\Rightarrow$ higher Ω_L^* / δ^* and τ^* increase / output of both types of firms will increase.

- (ii) $\Delta u > 0$: Ω_H^*, Ω_L^*, and δ^* do not change; only τ^* decreases - logical result: relatively more firms investing in cyberprotection implies lower investment by each of them to attain the optimal result. Output of L firms is maintained; output of H firms is also maintained (the decrease in τ^* is compensated by the increase in u and, according to (38), there is no change on the available protection and, thus, on output).

- (iii) $\Delta \iota > 0$: Ω_H^*, Ω_L^*, and δ^* do not change; only τ^* decreases - logical result: a lower degree of non-rivalry in selling protection implies H firms will invest more to keep more protection and to profit more from trading. Output does not change for any of the firms for reasons similar to those of the previous item.

- (iv) $\Delta \vartheta > 0$: l.h.s. of (37) does not shift; r.h.s. of (37) shifts right $\Rightarrow$ higher Ω_H^* (this is the positive effect that innovation from cybersecurity has over knowledge when H firms increase cybersecurity in response to cybercrime) / Ω_L^* remains unchanged / δ^* increases (L firms demand more security to face higher risks) / τ^* increases due to the increase on δ^* and directly on ϑ. Output levels will increase, given the corresponding expressions.

- (v) $\Delta b > 0$: l.h.s. of (37) does not shift; r.h.s. of (37) shifts right $\Rightarrow$ higher Ω_H^* / Ω_L^* remains unchanged / τ^* increases because of the increase in Ω_H^*; δ^* does not change / the output of L firms does not change / the output of H firms increases.

Appendix F Simulating the data economy: a numerical example

Take the values in Table 1.

Parameter	Symbol	Value
Data endowment	z	10
Coefficent of the AR(1) process	ρ	0.9
Variances	σ^2	2.5
Share of H-type firms	u	1/3
Non-rivalry parameter	ι	0.6
Intertemporal discount factor	β	0.96
Data risk index	ϑ	0.75
Maximum quality	$\overline{A}$	25
Inonvation externality	b	0.035
Capital cost	r	1

TABLE 1 - VALUES OF PARAMETERS.

For these values of parameters: $\Omega_H^* = 3.224$ and $\Omega_L^* = 1.609$. These results are found in the intersection of the l.h.s. and r.h.s. of (37) and (44) **[Fig.1]**

Applying the corresponding formulas, $\delta^* = 0.130$ and $\tau^* = 1.296$ (these are both positive values and, therefore, firms engage in cybersecurity investment and cybersecurity trading).

Replacing the equilibrium values in the expressions for output and profits, $A_H^* = 23.207$ and $A_L^* = 21.879$ ($A_H^* > A_L^*$); $\Pi_{H,\text{sec}}^* = 21.068$ and $\Pi_{L,\text{sec}}^* = 20.800$ ($\Pi_{H,\text{sec}}^* > \Pi_{L,\text{sec}}^*$). Also, $Y^* = uA_H^* + (1-u)A_L^* = 22.321$.

F.1 Comparative statics

How do steady state values change with data risk?

Recall that $\vartheta \in [0,1]$. Evaluating the model for different values of ϑ (and letting all other values be as in Table 1), we find two thresholds: at $\vartheta = \frac{z-\Xi_L}{z} = 0.6583$, optimal security purchasing, δ^*, changes from negative to positive, implying that firms L buy protection only for $\vartheta > 0.6583$. For $\vartheta \leq 0.6583$, H firms have to choose whether to invest in protection or not, knowing that they will sell no protection. They compare profits $\Pi_{H,\text{sec}}^*$ and $\Pi_{H,cy}^*$; these are equal around $\vartheta = 0.3$. For $\vartheta > 0.3$, H-type firms invest in protection, otherwise they do not.

Fig.2 draws profits without protection for both firms (these are identical), the profits of the H firms with security investment, and the profits of the L firms under security trading. The two mentioned thresholds are highlighted.

Hence: for $\vartheta \leq 0.3$, H-firms do not invest in cyberprotection and L-firms do not buy protection; for $0.3 < \vartheta \leq 0.6583$, H firms invest in protection and L firms buy no protection; for $\vartheta > 0.6583$, H-type firms invest in protection and L-type firms buy protection. In this last segment, the higher the value of ϑ, the more the H firms invest and the more L firms buy.

Fig. 3 presents the investment and trading levels. Again, the two thresholds are clear

(notice the second jump in τ^*; this occurs because to the right of that point, H-type firms need to invest in security for their one use but also to sell to firms in the L group).

Fig. 4: output of each type of firm and aggregate output, for different levels of data risk. In the first segment, the output is the same (the firms are identical); in the second segment, L firms face increasing risk but do not protect and, consequently, output falls (because the stock of knowledge falls); H firms start investing in cybersecurity what has the innovation side effect and, therefore, they are able to increase output. In the third segment, H firms continue to invest in cyberprotection and innovate; L firms start purchasing security that they cannot use to innovate but that prevents output from falling (i.e., it allows to maintain the stock of knowledge as the data risk increases).

The aggregate output is a weighted average of the output of the two types of firms (recall that, in the example, L firms are two thirds of the total number of firms). Concerning the evolution of Y^* as ϑ increases, one notices that an initial fall is counteracted when H firms start to invest in protection, and this process gains a new impetus when L firms start protecting as well.

We can recover the representation of profits in Fig.2, to draw the actual profits, given the choices of firms on whether to get protection or not. **Fig. 5** clarifies again the existence of three stages and the fact that cybercrime is much less harmful for H-type firms, because these make use of the innovation externality that cybersecurity allows for.

F.2 Does data growth cause economic growth?

In the model, there are various parameters whose values can change - $\vartheta, \iota, u, b, ...$ - but only one can grow in a sustained way over time, which is the endowment of data, z. The question is: if one makes z to increase over time at a constant rate, will the economy's output also grow over time at a constant rate?

The answer is no: simulations show that although the increase in z leads to increases in Ω_H^* and Ω_L^*, they also lead to falls in δ^* and τ^* (more data and a same data risk lead to the need of less protection). For large values of z, τ^* becomes zero, and without investment in cybersecurity there is no data risk induced innovation and the maximum quality of output cannot expand. The increases in Ω_H^* and Ω_L^* are associated with decreasing marginal returns and, therefore, although z might grow in a sustained way, this is not accompanied by an increase in the firms' output.