Review of Economic Studies (2022) **0**, 1–23 https://doi.org/10.1093/restud/rdac056

Advance access publication 13 August 2022

Data-intensive Innovation and the State: Evidence from AI Firms in China

MARTIN BERAJA
MIT and NBER
DAVID Y. YANG
Harvard University, NBER, and CIFAR
and
NOAM YUCHTMAN
LSE, CEPR, and CESifo

First version received February 2021*; Editorial decision January* 2022*; Accepted August* 2022 (*Eds.*)

Developing artificial intelligence (AI) technology requires data. In many domains, government data far exceed in magnitude and scope data collected by the private sector, and AI firms often gain access to such data when providing services to the state. We argue that such access can stimulate commercial AI innovation in part because data and trained algorithms are shareable across government and commercial uses. We gather comprehensive information on firms and public security procurement contracts in China's facial recognition AI industry. We quantify the data accessible through contracts by measuring public security agencies' capacity to collect surveillance video. Using a triple-differences strategy, we find that *data-rich* contracts, compared to *data-scarce* ones, lead recipient firms to develop significantly and substantially more commercial AI software. Our analysis suggests a contribution of government data to the rise of China's facial recognition AI firms, and that states' data collection and provision policies could shape AI innovation.

Key words: Data, Innovation, Artificial intelligence, China, Innovation policy, Privacy, Surveillance.

JEL Codes: O30, P00, E00, L5, L63, O25, O40

1. INTRODUCTION

Artificial intelligence and machine learning ("AI" for brevity) technologies hold the potential to transform the modern world. Developing AI is *data-intensive*. Up to now, economists have emphasized how data collected by private firms shapes the process of AI innovation (Agrawal, Gans and Goldfarb, 2019; Jones and Tonetti, 2020). Yet, throughout history and up to the present, states have also collected massive quantities of data (Scott, 1998). Because of states' dominant role in domains such as public security, health care, education, and basic science,

The editor in charge of this paper was Nicola Gennaioli.

government data collected in these areas exceed in magnitude and scope available data collected by the private sector, or may lack private substitutes altogether.

A common way in which private AI firms gain access to valuable government data is by providing services to the state. Consider the facial recognition AI industry in China—a leading AI sector in a country at the technological frontier.[1] In order to develop accurate facial recognition algorithms, firms in this sector require enormous amounts of training data—for example, video streams of faces from different angles. The public security units of the Chinese state collect precisely this form of data through their surveillance apparatus, and contract with AI firms to process such data. AI firms providing services to these public security units thus gain access to government surveillance data, which can be inputs into improved algorithms and thus innovation.

Importantly, innovation stimulated by government data can go well beyond the government sector. To the extent that government data or trained algorithms are *shareable*, they can be used to develop AI products for much larger commercial markets—for instance, facial recognition platforms for retail stores. Moreover, firms receiving access to government data may learn how to manage and productively utilize large datasets, another valuable input into commercial innovation. Therefore, receiving a procurement contract allowing access to government data may fuel commercial AI innovation, potentially overcoming the crowd-out of resources allocated to serving the state.

In this article, we ask: does access to government data when providing AI services to the state stimulate commercial AI innovation? We answer this question in the context of the facial recognition AI sector in China. We collect comprehensive data on AI public security contracts and AI firms' software production, and we classify AI procurement contracts as data-rich or data-scarce depending on the size of the local surveillance camera network. We find that the receipt of a data-rich contract differentially stimulates commercial AI software innovation. Our findings suggest that access to government data may have contributed to Chinese firms' emergence as leading innovators in facial recognition AI technology—indeed, this has coincided with the expansion of the Chinese government's procurement of AI and surveillance capacity (see Figure 1). More generally, our findings indicate a role for the state in data-intensive economies that goes beyond regulating privately collected data out of anti-trust or privacy concerns (e.g. Tirole, 2021; Aridor, Che, Nelson and Salz, 2020). States' AI procurement and policies of government data collection and provision could, whether intentionally or not, stimulate and shape AI innovation in a range of sectors.

We begin by presenting a simple partial equilibrium model of AI software production. There are two types of data—government and private—that are gross substitutes in the production of new AI software (i.e. product innovation). Importantly, government data can only be accessed through obtaining a contract to produce government AI for the state. Software production is also a function of other non-shareable inputs (such as labour) as well as sharable ones (such as data management software). These elements allow for the possibility of both crowding-in and crowding-out of commercial AI innovation when firms contract with the state. The model allows us to specify the parameter of interest in our estimation—the change in commercial software production resulting from a change in government data accessed by the firm—and clarifies the mechanisms through which government data could affect commercial innovation. Finally, the model highlights key threats to identification: in particular, firm characteristics correlated with receipt of (data-rich) contracts and productive inputs accessed through procurement contracts alongside government data.

[1] China is the world's largest producer of AI research (see the "China AI Development Report, 2018," available online at https://bit.ly/2IWAo7R. Facial recognition AI is among the top three AI technologies in terms of projected revenues (Perrault, Shoham, Brynjolfsson, Clark, Etchemendy, Grosz, Lyons, Manyika, Mishra and Niebles, 2019).

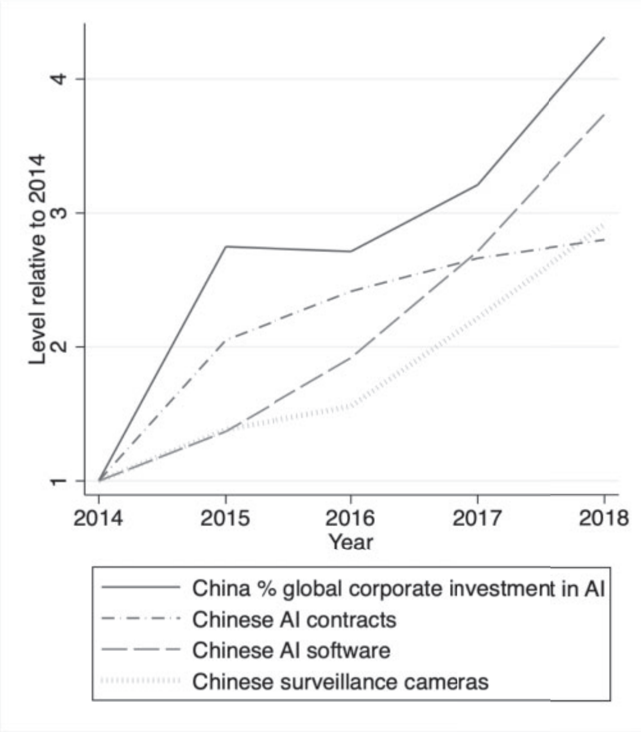

FIGURE 1

This figure displays four metrics relevant to China's AI development, relative to their levels in 2014. First, the percentage of global corporate investment in AI originating from China, sourced from NetBase Quid (initial value = 4%). Second, the number of facial recognition AI contracts procured by the Chinese government, sourced from the Chinese Government Procurement Database (initial value = 1899). Third, the cumulative amount of AI software produced by Chinese firms, sourced from the Ministry of Industry and Information Technology (initial value = 33,340). Fourth, the number of surveillance cameras procured by the Chinese government, sourced from the Chinese Government Procurement Database (initial value = 755,134).

Reflecting this model, our empirical strategy compares changes in firm software output following the receipt of *data-rich* versus *data-scarce* government contracts. In order to operationalize it, we overcome three data challenges. First, linking AI firms to government contracts. To do so, we collect data on (approximately) the universe of Chinese facial recognition AI firms and link this data to a separate database of Chinese government contracts, issued by all levels of the government. Second, quantifying AI firms' software production and, as important, classifying firms' software by intended use. We do this by compiling data on all Chinese facial recognition AI firms' software development based on the digital product registration records maintained by the Chinese government. Using a Recurrent Neural Network model, we categorize software products based on whether they are directed towards the commercial market or government use. Third, measuring the amount of government data to which AI firms receive access. To do this, we focus on contracts awarded by public security agencies to AI firms. Within this set of contracts, we measure the data provided by a contract using the agency's local surveillance network capacity to capture high-resolution video of faces on the streets: namely, the number of high-resolution surveillance cameras that had previously been purchased by government units in the public security agency's prefecture. We define a data-rich contract as one that came from a public security agency located in a prefecture with above-median surveillance capacity at the time the contract was awarded, whereas a data-scarce contract is one coming from a public security agency located in a prefecture with below-median surveillance capacity.

With these newly constructed datasets, we estimate our parameter of interest: the causal effect of access to government data on commercial AI innovation. We compare the post-contract increase in software releases of firms that receive data-rich and data-scarce public security contracts. This comparison allows us to hold fixed firms' (time-invariant *and* time-varying) selection into receipt of a public security contract. Moreover, by exploiting variation in data-richness within the set of public security contracts, it also allows us to pin down the importance of access to *government data* rather than other benefits of government contracts, such as capital, reputation, and political connections.[2] We find that receipt of a data-rich contract *differentially* increases commercial software production, relative to receipt of a data-scarce contract, by around 2 new software products over 3 years. Importantly, we identify a significant effect of data-rich contracts on government software production over the same period as well, indicating that the increase in commercial innovation overcomes the crowding out of non-shareable inputs used to produce government software.

We evaluate key threats to identification highlighted in the model. First, firm characteristics affecting software production may be correlated with receipt of (data-rich) public security contracts. We directly account for fixed firm characteristics correlated with receipt of a data-rich contract in our empirical specification by including firm fixed effects. While we cannot directly control for unobserved time-varying sources of selection into data-rich contracts, it is reassuring that our event-study estimates show no differential software production prior to receipt of a data-rich contract, suggesting software production among firms receiving data-rich and data-scarce contracts would have followed parallel trends. We can also directly account for the time-varying effects of selection into contracts based on firms' underlying productivity, as measured by their pre-contract characteristics such as software production, establishment year, and capitalization. Second, productive inputs may be accessed through data-rich procurement contracts alongside government data. We consider and rule out the following alternative mechanisms through which data-rich contracts may stimulate firms' commercial innovation: access to capital, distinct tasks required by data-rich contracts, reputational consequences, access to markets and associated commercial opportunities, and connections with the local governments.

Finally, we assess the contributions of direct and indirect mechanisms through which government data could affect commercial innovation. We first provide evidence of non-data shareable inputs arising from accessing greater quantities of government data: we find that production of non-AI, data-complementary software (e.g. software supporting data storage and transmission) significantly, and differentially, increases after firms receive data-rich public security contracts. We then evaluate whether such increase in shareable inputs could account for the entirety of the increased commercial innovation that we observe. We use pre-contract data-complementary software production as a summary statistic for firms' potential to benefit from the development of additional non-data shareable inputs arising from a data-rich contract. We find that controlling for this potential to benefit from additional non-data shareable inputs interacted with the full set of time period fixed effects barely changes our estimated effects of government data on commercial AI innovation. This suggests an important direct effect of government data and improved algorithms due to their being shareable across uses.

Our work most directly contributes to an emerging literature on the economics of AI and data (see e.g. Agrawal, Gans and Goldfarb, 2018; Acemoglu and Restrepo, 2019;

[2] Our empirical analysis is thus related to studies of the causal effects of government contracts on firm performance (e.g. Ferraz, Finan and Szerman, 2015) but considers variation *within* the set of firms receiving these contracts.

Aghion, Jones and Jones, 2019; Farboodi, Mihet, Philippon and Veldkamp, 2019). We add to this literature by examining the role of *government*-collected data and the direct and indirect ways in which it can shape commercial AI innovation.[3] We particularly highlight the shareablility of government data and trained algorithms across uses within firms, complementing Aghion *et al.* (2019) and Jones and Tonetti (2020) who study the non-rivalry of private data across firms.

We also contribute to the literature studying government policies that shape innovation (see Bloom, Van Reenen and Williams, 2019 for a review). Our work reveals that government data collection and provision to firms can act as an innovation policy, either intentionally or incidentally. Our work also indicates that government data collection and provision stimulate commercial innovation through mechanisms that share features with other government policies, from space exploration (Alic, Branscomb, Brooks and Carter, 1992; Azoulay, Fuchs, Goldstein and Kearney, 2019), to the internet (Greenstein, 2015), to military technology (Moretti, Steinwender and Van Reenen, 2019; Gross and Sampat, 2020).[4] Much like scientific ideas, government data can spur learning by doing and generate other intangible assets; in addition, we highlight that government data (and trained algorithms) themselves can be directly shared across uses, resulting in potentially faster and greater commercial spillovers. Empirically, we identify specific, causal mechanisms through which a shareable input affects commercial innovation at the firm level.

Finally, our work contributes to the literature studying the economic rise of China, joining a large literature that emphasizes the important role of the state (e.g. Lau, Qian and Roland, 2000; Brandt and Rawski, 2008; Song, Storesletten and Zilibotti, 2011). In highlighting the role of China's surveillance apparatus in commercial innovation, we contrast with a large literature attempting to explain China's spectacular growth despite its authoritarian institutions—for example, growth stimulated by competition for promotion (e.g. Li and Zhou, 2005; Jia, Kudamatsu and Seim, 2015), or bureaucratic rules of evaluation and rotation (Li, 2019).[5] We contribute to a nascent literature (e.g. Bai, Hsieh and Song, 2020) that identifies mechanisms through which China's autocratic power may actually promote economic growth.

In what follows, we present a simple conceptual framework in Section 2. Section 3 describes the empirical context and the data sources used for the analyses. Section 4 presents the main results. Section 5 concludes with a discussion of implications and direction of future work.

2. CONCEPTUAL FRAMEWORK

Consider a firm that produces AI software for both the state (government software) and the private sector (commercial software). Developing both types of software is data-intensive: it requires algorithms trained with data. There are two types of data: one collected by the state (government data) and one collected by the private sector (private data). As noted above, there exist important domains where government data far exceeds in magnitude and scope any private

[3] In so doing, our analysis complements a recent literature studying the effects of government data on other sectors. For example, Williams (2013) and Nagaraj (2021) study settings in which the non-excludability of government research—mapping the genome and mapping the Earth—shapes private sector outcomes in biotechnology and mineral extraction, respectively.

[4] Incidental industrial policy is also documented by Slavtchev and Wiederhold (2016) and Nagle (2019). Our finding of a within-firm spillover to products *other than* those contracted on contrasts with firms' tendency to specialize after a specific government demand shock, as seen in Clemens and Rogers (2020).

[5] On China's growth and innovation more specifically, see, among others, Song *et al.* (2011), Khandelwal, Schott and Wei (2013), Roberts, Yi Xu, Fan and Zhang (2017), Cheng, Jia, Li and Li (2019), Wei, Xie and Zhang (2017), and Bombardini, Li and Wang (2018). On various economic distortions caused by China's political institutions, see, among others, Chen, Ebenstein, Greenstone and Li (2013), Fisman and Wang (2015), and He, Wang and Zhang (2020).

substitutes (e.g. surveillance video from street cameras). Moreover, in many cases government data is not publicly available—a firm only gains access to such government data when obtaining a contract from the state to produce government software.

Formally, a firm that has obtained a contract produces commercial software q_c and government q_g software with the following technologies:

$$q_c = F_c\left(d_g, d_p, s, n_c; X, C\right)$$
$$q_g = F_g\left(d_g, d_p, s, n_g; X, C\right).$$

We denote by d_g the amount of government data provided by the contract and by d_p other private data inputs that the firm may have access to. Note that the same d_g and d_p enter the production of both types of software. This reflects the fact that government and private data (or the algorithms trained with them) are *shareable* across uses: they can be used to develop software for both government and commercial purposes (i.e. products targeting different customer types).[6]

Software production is also a function of other inputs s that are shareable across uses—such as data management software or firms' capacity and protocols to handle large datasets—as well as non-shareable inputs n_c and n_g—such as human and physical capital. Finally, we let firm characteristics X (e.g. the firm's underlying productivity), and government contract characteristics C other than the amount of government data (e.g. political connections provided by the contract) shape software production too.

Consider a comparison between two identical firms (fixed X) that have obtained contracts that only differ in the quantity of government data d_g made available to them but not in other characteristics (fixed C). Given a difference in government data Δd_g and fixed characteristics $\{\bar{X}, \bar{C}\}$, the difference in commercial software production $\Delta q_c\left(\cdot; \bar{X}, \bar{C}\right)$ is:

$$\frac{\Delta q_c\left(\cdot;\bar{X},\bar{C}\right)}{\Delta d_g} = \underbrace{\frac{\partial F_c\left(\cdot;\bar{X},\bar{C}\right)}{\partial d_g}}_{\text{Direct effect of data}} + \underbrace{\frac{\partial F_c\left(\cdot;\bar{X},\bar{C}\right)}{\partial d_p}\frac{\Delta d_p}{\Delta d_g} + \frac{\partial F_c\left(\cdot;\bar{X},\bar{C}\right)}{\partial s}\frac{\Delta s}{\Delta d_g} + \frac{\partial F_c\left(\cdot;\bar{X},\bar{C}\right)}{\partial n_c}\frac{\Delta n_c}{\Delta d_g}}_{\text{Indirect effect of data}}.$$

This will be the parameter of interest that we aim to estimate: the causal effect of government data on commercial software development. Note that it is composed of two elements: (i) a direct positive effect of government data that arises due to the shareability of government data (or algorithms); and (ii) an indirect effect which can amplify or dampen—or even reverse—the direct effect. The indirect effect will tend to augment the direct effect when other non-data shareable inputs s increase as well, for example, because the firm's capacity to manage and utilize datasets improves when provided access to more government data (i.e. a form of learning by doing). On the other hand, the indirect effect will tend to offset the direct effect when fulfilling the contract crowds out non-shareable inputs n_c from commercial software production to be used for government software production, or when private data d_p is substituted for by government data d_g.

The expression above reveals that obtaining a contract with greater access to government data can stimulate commercial innovation $\left(\frac{\Delta q_c\left(\cdot;\bar{X},\bar{C}\right)}{\Delta d_g} > 0\right)$ when the direct effect due to the shareability of government data and indirect effects arising from other shareable inputs are strong. However,

[6] Technological or legal reasons may limit the extent to which government data is shareable across uses. Yet, if the algorithms trained with such data can be shared and used for producing commercial software, then access to government data would stimulate commercial innovation through similar mechanisms. For the purposes of this article, data or trained algorithms being shareable are indistinguishable.

the total effect could be nil when government and private data are sufficiently substitutable (and other non-data shareable inputs do not change), or even negative when the crowding-out of non-shareable inputs due to government software production is sufficiently strong.

This thought experiment illustrates the empirical approach we follow to estimate our parameter of interest: comparing the changes in commercial software output between firms that obtained *data-rich* versus *data-scarce* contracts. It also reveals the two main threats to identification our empirical work will need to account for: (i) firms obtaining data-rich contracts may have different characteristics from those obtaining data-scarce contracts (X differs); and (ii) data-rich contracts may differ from data-scarce contracts along dimensions other than the amount of government data they provide (C differs). When X or C differ alongside d_g, the comparison between firms would not deliver the parameter of interest but also incorporate the effects of these other confounding factors.

3. THE STATE AND CHINA'S FACIAL RECOGNITION AI INDUSTRY

3.1. *Empirical context*

China's facial recognition AI sector is a prototypical setting in which to examine the impact of access to government data on commercial innovation. First, because facial recognition AI is extremely data-intensive: the development of the technology requires access to large datasets containing faces. Second, public security units of the Chinese state contract with facial recognition AI firms to provide them services in order to monitor citizens. Third, because these units collect huge amounts of surveillance data that firms can gain access to when obtaining a contract. Indeed, the value of government data is clear to private sector entrepreneurs: in 2019, a founder of a leading Chinese AI firm stated, "The core reason why [Chinese] AI achieves such tremendous success is due to data availability and related technology. Government data are the biggest source of data for AI firms like us."[7] Importantly, data acquired privately are not currently a close substitute for government data: in 2019, the former premier, Li Keqiang, stated that, "At this time, 80% of the data in China is controlled by various government agencies".[8]

Applying our conceptual framework to this context, consider an example in which a private firm receives a procurement contract to provide facial recognition software and data analysis services to a municipal police department in China. The firm implicitly receives access to large quantities of government data which are not publicly available. Such data include video from street surveillance cameras, and, potentially, labelled images with names and faces of individuals. The firm uses this data to train an AI algorithm; for example, a "tracking" algorithm that matches faces across video feeds or a "detection" algorithm that matches faces from video to the database of individuals. Then, the government data (or a base algorithm trained with it) can used to produce a separate trained algorithm that results in a commercial AI product; for example, AI software designed for retail firms that may wish to track or detect individual shoppers throughout their stores, and then predict their consumption choices.

3.2. *Data sources*

Operationalizing our empirical analysis faces three data-related empirical challenges: first, the need to link AI firms to government contracts; second, the need to compile information on AI

[7] *Source*: Chinese People's Political Consultative Conference, https://bit.ly/3gdo2T6.

[8] *Ibid*. It is important to note that Chinese government support of AI innovation is not limited to data provision but also includes a range of subsidies. Industrial policy that broadly affects all firms (whether or not they receive government data) is thus an important characteristic of the setting we study. It is also more broadly a characteristic of AI innovation around the world.

firms' software production, and specifically whether a given software is intended for commercial or other uses (e.g. for government use); and, third, the need to measure the quantity of government data to which firms have access. We address these challenges by constructing a novel dataset combining information on Chinese facial recognition AI firms and their software releases, and information on local governments' procurement of AI software and of surveillance cameras.[9]

Linking Chinese facial recognition AI firms to government contracts We identify (close to) all active firms based in China producing facial recognition AI using information from *Tianyancha*, a comprehensive database on Chinese firms that draws information from official, public records.[10] We extract firms that are categorized as facial recognition AI producers by the database, and we validate the categorization by manually coding firms based on their descriptions and product lists. We complement the *Tianyancha* database with information from *Pitchbook*, a database owned by Morningstar on firms and private capital markets around the world.[11] Using the overlap between sources, we validate the coding of firms identified in the *Tianyancha* database. We also supplement the *Tianyancha* data by adding a small number of AI firms that are listed by *Pitchbook* but omitted by *Tianyancha*. Overall, we identify 7837 Chinese facial recognition AI firms.[12] We also collect an array of firm level characteristics such as founding year, capitalization, major external financing sources, as well as subsidiary and mother firm information.

We extract information on 2,997,105 procurement contracts issued by all levels of the Chinese government between 2013 and 2019 from the Chinese Government Procurement Database, maintained by China's Ministry of Finance.[13] The contract database contains information on the good or service procured, the date of the contract, the monetary size of the contract, the winning bid, as well as, for a subset of the contracts, information on bids that did not win the contract.

We focus on contracts awarded by public security agencies to AI firms to analyse data drawn from local surveillance networks. These contracts provide firms with access to massive quantities of data, collected for monitoring purposes. Take, as an example from our dataset, a public security contract signed between an AI firm and a municipal police department in Heilongjiang Province to "increase the capacity of its identity information collection system" on 29th August 2018. The contract specifies that the AI firm shall provide a facial recognition system that can store and analyse at least *30 million* facial images—a substantial amount of data to which the firm obtains access.

We begin with a comprehensive set of public security agency procurement contracts, including 410,510 contracts in total. Within this set of public security contracts, we focus on the ones issued by prefecture level governments. This includes the following four types of public security contracts from the Chinese Government Procurement Database: (i) all contracts for China's flagship surveillance/monitoring projects — *Skynet Project*, *Peaceful City Project*, and *Bright Transparency Project*; (ii) all contracts with local police departments; (iii) all contracts with the border control and national security units; and, (iv) all contracts with the administrative units for domestic security and stability maintenance, the government's political and legal affairs commission, and various "smart city" and digital urban management units of the government.

[9] Supplementary Appendix Table A.1 describes the core variables and their sources.

[10] For example, a primary source of firms' information compiled by Tianyancha is the National Enterprise Credit Information Publicity System, maintained by China's State Administration for Industry and Commerce. See Supplementary Appendix Figure A.1 for an example entry.

[11] See Supplementary Appendix Figure A.2 for an example entry.

[12] These firms fall into three categories: (i) firms specialized in facial recognition AI (e.g. Yitu); (ii) hardware firms that devote substantial resources to develop AI software (e.g. Hik-Vision); and (iii) a small number of distinct AI units within large tech conglomerates (e.g. Baidu AI).

[13] See Supplementary Appendix Figure A.3 for an example contract.

To identify public security contracts procuring facial recognition AI, we match the contracts with the list of facial recognition AI firms, identifying 28,023 procurement contracts involving at least one facial recognition AI firm.[14] Many firms receive multiple contracts; overall, 1095 facial recognition AI firms in our dataset receive at least one contract.

Counting and classifying novel facial recognition AI software products We collect all software registration records for our facial recognition AI firms from China's Ministry of Industry and Information Technology, with which Chinese firms are required to register new software releases and major upgrades. We are able to validate our measure of software releases (using a single large firm), by cross-checking our data against the IPO Prospectus of MegVii, the world's first facial recognition AI company to file for an IPO.[15] We find that our records' coverage is comprehensive (at least in the case of MegVii): MegVii's IPO Prospectus contains 103 software releases, all of which are included in our dataset.

The count of new software releases (and major upgrades) represents *product innovation*.[16] While we are unable to observe firms' profitability, we observe that facial recognition AI firms that develop more software have significantly and substantially higher market capitalization, reflecting the economic value of such innovation (see Supplementary Appendix Figure A.5). In addition to quantity, we discuss measures of the quality of product development through the release of facial recognition AI software that involves video, a sophisticated and data-demanding facial recognition application (see Section 4.2).

We use a recurrent neural network (RNN) model with tensorflow—a frontier method for analysing text using machine learning—to categorize software products according to their intended customers and (independently) by their function. Our categorization by customer distinguishes between software products developed for the government (e.g. "smart city—real time monitoring system on main traffic routes") and software products developed for commercial applications (e.g. "visual recognition system for smart retail"). We allow for a residual category of general application software whose description does not clearly specify the intended user (e.g. "a synchronization method for multi-view cameras based on FPGA chips"). By coding as "commercial" only those products that are specifically linked to commercial applications, and excluding products with ambiguous use, we aim to be conservative in our measure of commercial software products.

Our categorization by function first identifies software products that are directly related to AI (e.g. "a method for pedestrian counting at crossroads based on multi-view cameras system in complicated situations"). Within the category of AI software, we also separately identify a subcategory of software that is particularly data-intensive: video-based facial recognition, which (as opposed to static images) requires N-to-1 or even N-to-N matching algorithms that are extremely data demanding. Finally, we identify a separate category of non-AI software products that are data-complementary, involving data storage, data transmission, or data management (e.g. "a computer cluster for webcam monitoring data storage").

To implement the two dimensions of categorization using the RNN model, we manually label 13,000 software products to produce a training corpus. We then use word-embedding to

[14] We present the cumulative number of AI procurement contracts in Supplementary Appendix Figure A.4 (top panel), as well as the flow of new contracts signed in each month (bottom panel). Both public security and non-public security AI contracts have steadily increased since 2013.

[15] Source: Hong Kong Stock Exchange, https://go.aws/37GbAZG.

[16] The National Science Foundation defines product innovation as "the market introduction of a new or significantly improved good or service with respect to its capabilities, user-friendliness, components, or subsystems" in its Business Enterprise Research and Development Survey (see https://www.nsf.gov/statistics/srvyberd/). See also Bloom, Jones, Van Reenen and Webb (2020).

convert sentences in the software descriptions into vectors based on word frequencies, where we use words from the full dataset as the dictionary. We use a Long Short-Term Memory (LSTM) algorithm, configured with 2 layers of 32 nodes. We use 90% of the data for algorithm training, while 10% is retained for validation. We run 10,000 training cycles for gradient descent on the accuracy loss function. The categorizations perform well in general: we are able to achieve 72% median accuracy in categorizing software customer and 98% median accuracy in categorizing software function in the validation data. Supplementary Appendix Figure A.6 shows the summary statistics of the categorization output by customers and by function; and, Supplementary Appendix Figure A.7 presents the confusion matrix (Type-I and Type-II errors) of the predictions relative to categorization done by humans.[17]

Measuring the quantity of government data to which firms have access Within the set of public security AI contracts, we identify those that are likely to be especially rich in data for facial recognition AI firms. We measure the data provided by a contract using the public security agency's local surveillance network capacity to capture video of faces on the streets in high-resolution: that is, the number of high-resolution surveillance cameras that had previously been purchased by government units in the agency's prefecture. This thus captures the amount of *identifiable* facial data that a facial recognition AI firm may gain access to.[18] Specifically, using 5837 prefectural government contracts for purchases of surveillance cameras, we sum the number of cameras procured in each prefecture up to a certain date and divide this by the prefecture's population to form a time-varying measure of the video surveillance capacity of a particular prefecture.[19] We measure data-richness using the density of cameras per capita because it proxies for the surveillance network's ability to observe the same faces multiple times, a key component of training data quality from a machine learning perspective. In a robustness specification below, we instead consider the counts of cameras.

Our empirical definition of a data-rich contract is one with a public security agency located in a prefecture that has above-median surveillance capacity (measured by cameras per capita) at the time the contract was awarded. Figure 2 shows the distribution of data-rich and data-scarce contracts across prefectures according to our definition.[20] We compare the effects of these data-rich public security contracts to data-scarce public security contracts, where data-scarce contracts are defined as those awarded by a public security agency located in a prefecture that has below-median surveillance capacity at the time the contract was awarded.

Summary statistics Table 1 presents summary statistics describing the firms in our sample. Firms receiving different types of contracts differ substantially from each other, so accounting for

[17] Supplementary Appendix Table A.2 presents the top words (in terms of frequency) used for the categorization. Supplementary Appendix Figure A.8 presents the density plots of the algorithm's category predictions. The algorithm is very accurate in categorizing software for government purposes. The algorithm is relatively conservative in categorizing software products for commercial customers and relatively aggressive in categorizing them as general purpose. In setting our categorization threshold for commercial software, we again aim to be conservative in our measure of commercial software products.

[18] Note that the existence of a national ID system in China likely implies that there may be limited variation across local public security agencies in *identified* personal images. Moreover, even if firms did not gain access to identified data, surveillance video alone would still be useful for many AI applications.

[19] This measure captures the stock of *newer* surveillance cameras at the time but not the older ones. The focus on newer cameras is appropriate given their higher resolution and thus greater usefulness in identifying and matching faces (see the Chinese government's directive on video surveillance: https://bit.ly/3dqdjU0). There are on average 77 surveillance camera contracts per prefecture. In Supplementary Appendix Figure A.9, we present a time series plot of the number of cameras in our data over time.

[20] By measuring data-richness at the time of the contract, we ensure that secular trends in surveillance capacity do not skew our measure toward coding later contracts as data-richer.

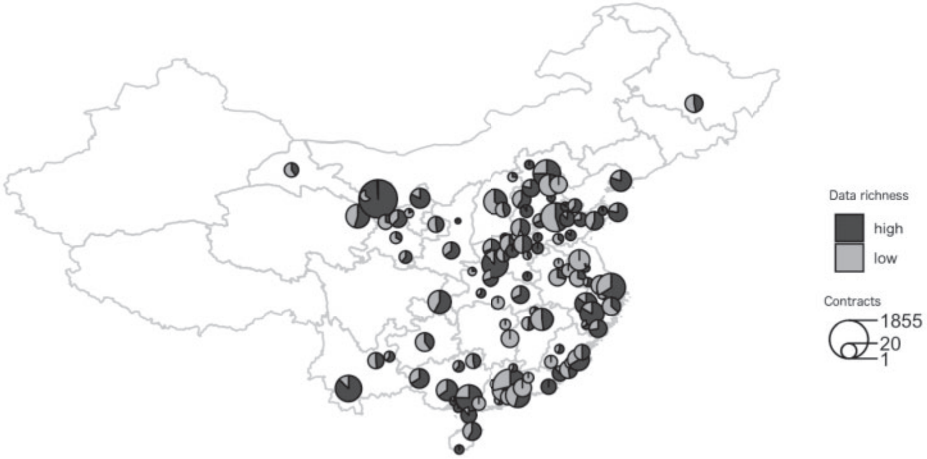

FIGURE 2

This figure illustrates public security AI procurement across China. Circle size indicates the number of first AI contracts awarded in the prefecture. Circle shading indicates the fraction of first AI contracts that were data-rich or data-scarce, where the within-prefecture variation comes from changes in the number of surveillance cameras over time.

differences (both observable and unobservable) between the firms receiving data-rich and data-scarce contracts will be crucial to identify the effects of the contracts. Supplementary Appendix Table A.3 presents summary statistics describing the contracts procuring AI services in our sample.[21] Data-scarce and data-rich contracts differ on dimensions other than in the quantity of data to which firms receive access, so accounting for alternative mechanisms (other than data provision) through which data-rich contracts might affect software production will be crucial to identifying the causal effects of interest.

4. EMPIRICAL ANALYSES

4.1. *Empirical model and identification strategy*

Our parameter of interest is the change in commercial AI software production resulting from a change in government data that the firm has access to through providing services to the state. We use a triple differences design to identify the effects of accessing government data on facial recognition AI firms' subsequent product development. The empirical strategy exploits variation across time and across firms in the receipt of a public security contract, and across the data-richness of the contracts that firms receive. Specifically, as in an event study design, we compare firms' AI software releases before and after they receive their first public security contracts, controlling for firm and time period fixed effects. To help pin down the importance of access to *government data*, rather than other benefits of government contracts, such as capital, reputation, and political connections, we in addition exploit variation in the data-richness (i.e. surveillance capacity) of the local public security agencies that issue the contracts.

[21] In Supplementary Appendix Table A.4, we provide descriptive statistics for the prefectures where contracts were issued, again disaggregating by the type of agency and by surveillance capacity.

TABLE 1
Summary statistics—firms and their production

	Any contract		Public security contract		Public security contract by surveillance capacity	
	Yes	No	Yes	No	High	Low
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Firm characteristics						
Year firm established	2009.3	2013.8	2008.9	2011.4	2007.5	2010.0
	(6.4)	(4.2)	(6.4)	(6.1)	(7.0)	(5.7)
Capitalization (millions USD)	22.8	5.1	26.4	4.1	35.3	19.9
	(210.3)	(42.8)	(229.1)	(14.4)	(295.0)	(165.4)
Rounds of investment funding	0.9	0.5	1.0	0.3	1.0	1.0
	(1.7)	(1.9)	(1.8)	(0.8)	(1.8)	(1.7)
Observations	1093	6041	919	174	387	532
Panel B: Software production before first contract receipt						
Total amount of software	22.7	14.6	23.8	14.8	27.4	21.2
	(37.9)	(24.5)	(39.9)	(16.4)	(45.0)	(35.8)
Commercial	9.0	6.3	9.4	6.7	10.1	8.8
	(17.1)	(12.5)	(17.9)	(9.6)	(20.1)	(16.1)
Government	7.3	4.0	7.8	4.1	10.0	6.3
	(16.3)	(8.2)	(17.2)	(7.0)	(17.7)	(16.6)
AI (video)	1.6	1.0	1.6	1.4	2.0	1.3
	(3.8)	(2.8)	(3.9)	(3.2)	(4.9)	(3.0)
Data-complementary	9.2	5.6	9.7	5.9	11.3	8.6
	(16.7)	(10.8)	(17.5)	(8.4)	(19.4)	(16.0)
Observations	956	6042	835	121	345	490

Notes: Variables in Panel A come from *Tianyancha*; variables in Panel B come from the Ministry of China's Ministry of Industry and Information Technology. "Total amount of AI software" is classified by sector (commercial or government), and by function (AI-video). Data-complementary software is distinct from AI software. Observations at the firm level. Standard deviations are reported below the means. Columns 1 and 2 split the firms into those receiving any government contract or not. For firms not receiving any contract, Panel B describes all software production during the entire sample period. Conditional on receiving at least one government contract, Columns 3 and 4 split the firms into those whose first contract is awarded by a public security agency. Conditional on receiving the first government contract from a public security agency, Columns 5 and 6 split the firms depending on whether their first public security contract is awarded by prefectures with high or low levels of surveillance capacity.

We test whether firms receiving data-rich public security contracts differentially increase their commercial software production following receipt of the contract. To do so, we estimate the following empirical model:

$$y_{it} = \sum_T \beta_{1T} T_{it} Data_i + \sum_T \beta_{2T} T_{it} + \alpha_t + \gamma_i + \epsilon_{it}.$$

The outcome variable, y_{it}, is the cumulative number of commercial software releases by firm i up to the 6-month period t. The explanatory variables of interest are the interaction terms between a set of dummy variables, T_{it}, indicating 6-month time periods before or since firm i received its first contract, and $Data_i$, a dummy variable indicating whether the firm's first contract was data rich, as defined above.[22] We also include a full set of firm fixed effects, γ_i, and time period fixed effects, α_t, in all specifications. We allow the error term ϵ_{it} to be correlated not only across observations for a single firm but also across observations for firms that are related by common

[22] We focus on the effect of the initial contract because the receipt of subsequent contracts is endogenous to firms' performance in their initial contracts—therefore being part of the *total effect* one would wish to capture.

ownership by a single mother firm. We cluster standard errors at the mother firm-level to be conservative.

The coefficients $\beta_2 T$ describe the software production of a firm around the time when it receives its first data-scarce public security contract; the sums of coefficients $\beta_1 T + \beta_2 T$ describe the software production of a firm around the time when it receives its first data-rich public security contract. The coefficients on the interaction terms, $\beta_1 T$, thus non-parametrically capture a firm's differential production of new software approaching or following the arrival of initial data-rich contracts, relative to data-scarce ones. The $\beta_1 T$ coefficients correspond to our parameter of interest, to the extent that we are able to account for confounding factors such as firms' selection into data-rich contracts and contract characteristics unrelated to government data.

Our empirical specification allows us to account for a range of such factors. By including time fixed effects, we account for time-varying sources of variation in software production common to all facial recognition firms (for example, government policies promoting AI). We are also able to address a range of concerns regarding firms' selection into procurement contracts. Note that our triple differences design does not require exogenous assignment of *all* contracts—in fact, we fully account for selection into contracts with public security agencies by examining variation within the set of firms that receive public security contracts. Our identifying assumption is the exogenous access to greater amounts of government data conditional on the receipt of any public security contract and the controls we include. We account for the time-invariant sources of selection into data-rich public security contracts by including firm fixed effects. One remains concerned about time-varying sources of selection into data-rich contracts, which we assess by examining pre-contract levels and trends of software production, and we further address in a robustness specification below by controlling for the time-varying effects of firms' pre-contract characteristics ($\sum_T T_{it} X_i$). One also may be concerned that data-rich contracts differ from data-scarce contracts along other dimensions than data that could shape software production, which we address in additional robustness specifications by controlling for the time-varying effects of several salient contract characteristics ($\sum_T T_{it} C_i$).

4.2. *Baseline estimates of the parameter of interest*

We begin our empirical analyses by estimating our baseline specification described in Section 4.1, comparing the effects of public security contracts in prefectures with above-median surveillance capacity (data-rich contracts) with those that have below-median surveillance capacity (data-scarce contracts). In Figure 3A, we plot the coefficients β_{1T} and their 95% confidence intervals, describing the *differential* cumulative commercial software production around the time when a data-rich public security contract was received, relative to a data-scarce public security contract (all coefficients are presented in Table 2, Column 1).

We find that the receipt of a data-rich public security contract is associated with differentially more commercial software production than receipt of a data-scarce public security contract: around 1.0 additional software products 1 year after the contract receipt, increasing to around 1.9 additional software products over a period of 3 years after the contract. Over the 3-year period, this represents an increase in commercial software production of 20.2% relative to the pre-contract level. While we discuss threats to identification in detail below, we note that the absence of pre-contract differences in software production levels or trends suggests a causal effect of a data-rich contract.

This increase in commercial software takes place alongside an increase in government software production. We estimate our baseline model but instead considering government software production as an outcome, and we present the results in Figure 3B (all coefficients are presented in Table 2, Column 2). We find that data-rich public security contracts generate 2.9 additional

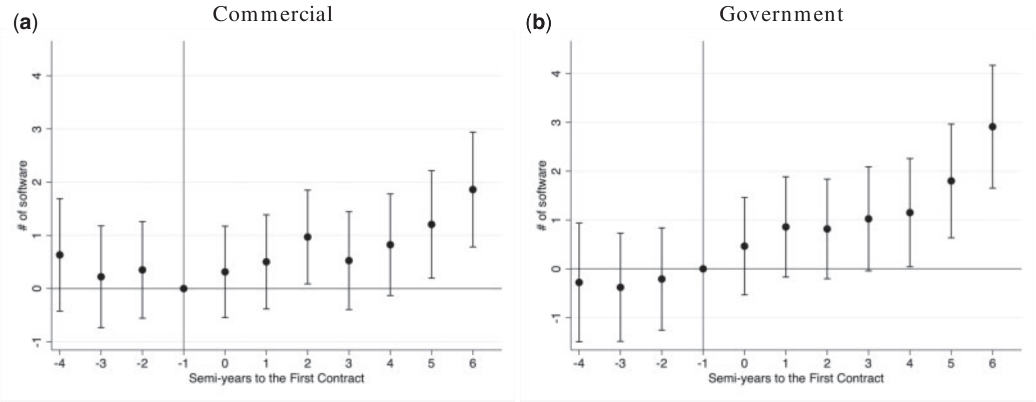

FIGURE 3

Differential cumulative software releases intended for commercial (left), for government uses (right), resulting from data-rich public security contracts, relative to data-scarce ones, controlling for firm and time period fixed effects. Data-rich contracts are defined as public security contracts in prefectures with above median surveillance capacity.

government software products (an increase by 51.9%) over 3 years after the receipt of the contract, compared to data-scarce contracts. Again, we find no pre-contract differences in levels or trends in government software production. Thus, the increase in commercial software production takes place despite the need to allocate resources to increase government software production.

Importantly, our results on commercial software indeed represent a differential increase in software production, rather than differential crowd-out. We observe an overall positive effect of both data-scarce and data-rich contracts on commercial software production, and differentially larger effects for the latter. We document this in Supplementary Appendix Figure A.10, which plots the coefficients β_{2T} and $\beta_{1T}+\beta_{2T}$ for commercial software production when a data-scarce and a data-rich public security contract are received, respectively.[23]

Robustness. Given the complex process of constructing our dataset, it is important to note that our findings are robust to varying several salient dimensions of our analysis (see Figure 4).

We begin by assessing the robustness of our results to variation in specifying our outcome of interest—measures of commercial software innovation. First, we restrict attention only to firms' new software releases (i.e. version 1.0) and major upgrades with a change in the first digit of the release number (i.e. versions 2.0, 3.0, etc.). Our baseline estimates remain largely unchanged, indicating that our results are not driven by minor software updates (see Panel A). An even more demanding check is to restrict attention to software that involves video—the most data demanding form of facial recognition AI. Indeed, we find significantly greater video facial recognition AI software production following receipt of a data-rich contract (see Panel B).

Second, we consider the three key parameters of choice in the RNN algorithm that we use to categorize software—timestep, embedding, and nodes. We vary these three parameters, re-configure the RNN LSTM algorithm, re-categorize software, and re-estimate the baseline empirical specification. We find that these algorithm parameter choices have no impact on our results (see Panel C). Third, we restrict attention to commercial software that we can classify with a very high degree of confidence by adjusting the LSTM classification threshold. The baseline

[23] The figure also shows an increase in government software production following the receipt of both data-rich and data-scarce public security contracts.

TABLE 2
Regression coefficients

	Commercial	Government
	(1)	(2)
4 semiyears before	−0.239	−0.177
	(0.231)	(0.268)
3 semiyears before	−0.180	−0.040
	(0.228)	(0.264)
2 semiyears before	−0.202	−0.002
	(0.225)	(0.261)
Receiving first contract	0.868***	0.750***
	(0.239)	(0.279)
1 semiyear after	1.663***	1.443***
	(0.250)	(0.289)
2 semiyears after	2.219***	2.243***
	(0.258)	(0.301)
3 semiyears after	3.122***	2.986***
	(0.287)	(0.334)
4 semiyears after	4.017***	3.984***
	(0.309)	(0.360)
5 semiyears after	4.857***	4.849***
	(0.337)	(0.389)
6 semiyears after	5.811***	5.595***
	(0.378)	(0.444)
4 semiyears before × data-rich	0.633	−0.279
	(0.539)	(0.620)
3 semiyears before × data-rich	0.222	−0.379
	(0.488)	(0.565)
2 semiyears before × data-rich	0.351	−0.209
	(0.463)	(0.535)
Receiving first contract × data-rich	0.314	0.465
	(0.438)	(0.508)
1 semiyear after × data-rich	0.502	0.858
	(0.451)	(0.524)
2 semiyears after × data-rich	0.969**	0.817
	(0.449)	(0.520)
3 semiyears after × data-rich	0.526	1.023*
	(0.470)	(0.544)
4 semiyears after × data-rich	0.823*	1.151**
	(0.487)	(0.565)
5 semiyears before × data-rich	1.205**	1.800***
	(0.515)	(0.594)
6 semiyears after × data-rich	1.861***	2.911***
	(0.550)	(0.642)

Notes: All regressions estimated on the sample of firms receiving first contracts from public security agencies. Baseline specification controls for time period fixed effects and firm fixed effects. * significant at 10% ** significant at 5% *** significant at 1%.

specification sets the threshold as 50%. We re-categorize software using higher classification thresholds of 60% and 70%, and these adjustments have no impact on our results (see Panel D).

We then assess the robustness of our results with respect to our definition of data-rich procurement contracts. First, we define data-richness of the contracts based on the absolute count of surveillance cameras, rather than the count per capita as used in the baseline specification. One can see that our results are unaffected by the modified definition (see Panel E). Second, we adjust our classification of (data-rich) public security contracts to exclude any ambiguous government agencies (e.g. contracts with the government headquarters, and smart city management and administrative bureaux could be meant to provide security services just for the government office building). This, too, has no impact on our results (see Panel F). Third, we consider an alternative

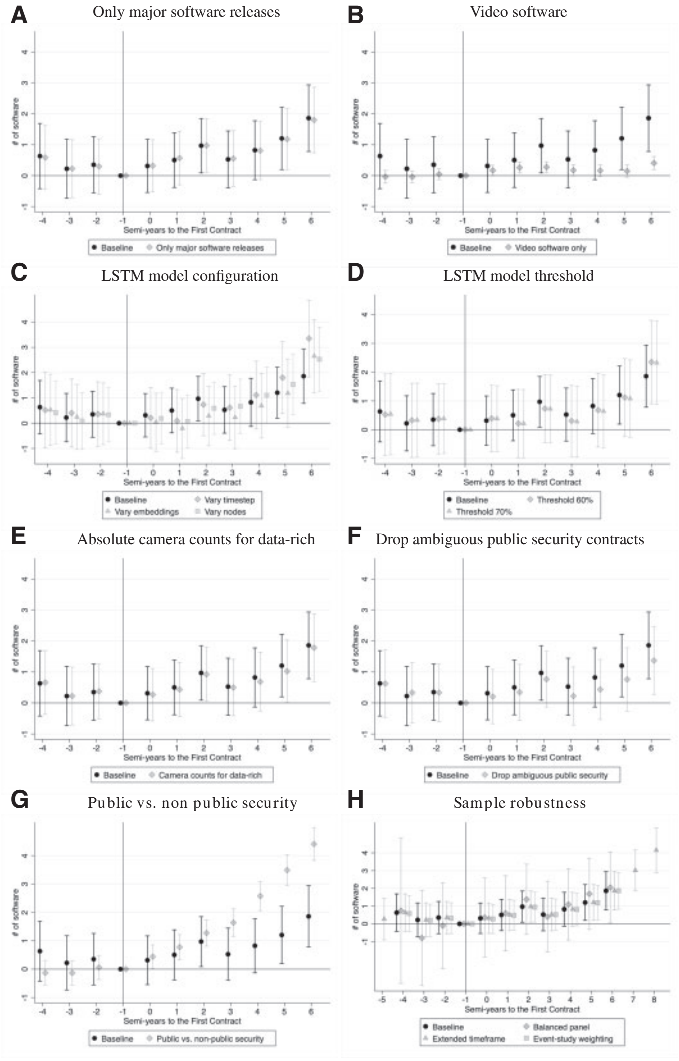

FIGURE 4

Panels replicate results from Figure 3(A) (plotted in circles). (A) adds results restricting software to only major releases (version X.0). (B) presents results with AI video software as the outcome. (C) varies the LSTM categorization model configuration. The diamonds show results for a LSTM model trained with a timestep of 10 instead (baseline level = 20), the triangles show results for a model trained with 16 embedding size instead (baseline level = 32), and the squares show results for a model trained with 16 nodes instead (baseline level = 32). (D) varies the LSTM categorization model's confidence threshold. The diamonds and triangles use thresholds of 60% and 70%, respectively (baseline level = 50%). (E)'s diamonds use above median absolute camera counts to define data-rich contracts (baseline specification uses cameras per capita). (F)'s diamonds exclude companies whose first contract may be with an ambiguous entity, or one that contains the keywords "local government" or "government offices" which may be used for either public security or non-public security. (G)'s diamonds show results based on an alternative definition of data-richness (public security contracts are classified as data-rich, while non-public security ones are data-scarce). (H) explores robustness with respect to sample, where the diamonds present results with a balanced panel, the triangles present results for an extended time frame (−5 quarters to 8 quarters after contract), and the squares present results over-weighting firms receiving no contract by a factor of 1000.

empirical definition of data-richness of government contracts. Procurement contracts awarded by a public security agency (even in locations with relatively few surveillance cameras) are most likely to provide access to massive, linkable, personal data, collected for monitoring purposes, while contracts with other, non-public security agencies likely provide access to less data.[24] We define a data-rich contract as one that came from a public security agency, whereas a data-scarce contract is one that did not. We re-estimate the baseline specification with this alternative definition of data-richness. The results are qualitatively unchanged (see Panel G). This analysis has the drawback of comparing the effects of types of contracts into which firm selection may differ substantially. However, when we examine the *direction* of selection into public security contracts (relative to non-public security ones), we find that it is often the *opposite* of what we observe when examining selection into data-rich public security contracts (relative to data-scarce public security contracts).[25] Finding the same qualitative effects using this alternative definition of data-richness argues against concerns that our results are driven by selection into data-richer contracts.

We next vary the sample used to estimate the baseline model. We consider a balanced panel of firms; an expanded window of time around the receipt of the first contract; and we address potential negative weighting issues in event studies by over-weighting firms receiving no contract by a factor of 1000 (Borusyak, Jaravel and Spiess, 2017). These changes do not affect our findings (see Panel H).

4.3. *Evaluating alternative hypotheses*

Motivated by our conceptual framework, we now evaluate whether the increased commercial software production observed following receipt of data-rich procurement contracts may be attributable to firms' characteristics related to selection into data-rich contracts, or characteristics of the contracts other than access to government data.

Firms' selection into data-rich contracts. Given the value of government contracts and government data, one naturally expects the sorting of firms into government contracts, in particularly the data-rich ones. Indeed, examining Table 1, Panels A and B, one observes that firms receiving data-rich public security contracts exhibit characteristics plausibly associated with higher underlying productivity: they tend to be older, better capitalized, and to have already produced more software prior to the receipt of first contract.[26]

Importantly, our empirical specification already accounts for important dimensions of sorting, including *any* form of sorting into public security contracts as well as sorting into data-rich contracts on time-invariant firm characteristics, whether observable or unobserved. Additional results suggest that sorting on time-varying characteristics cannot account for the effects of data-rich contracts that we observe. First, as noted above, we find no evidence of pre-contract differences in software production levels or trends, which one would expect if firms selected into data-rich government contracts as a function of their productivity trends. As another check, we

[24] Non-public security agencies (e.g. banks or schools) do not have access to large scale surveillance camera networks and cover narrower groups of individuals.

[25] For example, firms receiving public security contracts are better capitalized than firms receiving non-public security contracts (40 vs. 13 million USD; see Table 1), but firms receiving public security contracts in high-surveillance prefectures are less well capitalized than firms receiving public security contracts in low-surveillance prefectures (13 vs. 61 million USD).

[26] Consistent with selection, we also find that firms submit lower bids for data-rich contracts, and more firms submit bids for data-rich contracts (see Supplementary Appendix Figure A.11).

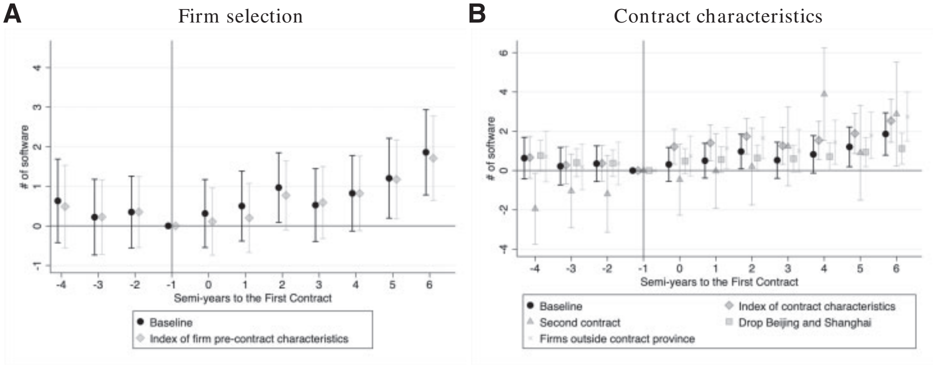

FIGURE 5

Panels replicate results from Figure 3A (plotted in circles). (A) adds results (plotted in diamonds) controlling for the time-varying effects of an index of firms' underling productivity (an inverse covariance weighted z-score of firms' establishment year, pre-contract capitalization, rounds of external financing prior to their first procurement contract, and total pre-contract software production). (B) adds results (plotted in diamonds) controlling for the time-varying effects of an index of contract characteristics (an inverse covariance weighted z-score of contract location GDP, tasks specified by the contract text, and contract monetary size). (B) also adds results (plotted in triangles) for the second contract received within the parent organization; results (plotted in squares) excluding contracts from Beijing and Shanghai; and results (plotted in x's) restricting the sample to firms that have their first contract outside of their home province.

account for firms' unobserved potential for productivity growth following contracts' receipt, by controlling for the time-varying effects of firms' underlying productivity. We proxy for firms' underlying productivity using their establishment year, pre-contract capitalization, rounds of external financing prior to their first procurement contract, and total pre-contract software production. We construct an index of firms' underlying productivity combining these proxies,[27] and we control for the time-varying effects of this index. Formally, we interact this index with a full set of time period fixed effects ($\sum_T T_{it} X_i$). As presented in Figure 5A, we find that these controls do not qualitatively or quantitatively affect our baseline estimates.[28]

Contract features other than government data. Procurement contracts that provide greater access to government data may also provide firms with a range of additional productive benefits. One basic consideration is that contracts may affect firms' software production through the provision of capital. Another possibility is that high-surveillance prefectures may also be richer; if so, a data-rich contract may stimulate additional software production by providing access to a richer commercial market. Yet another possibility is that data-rich contracts may require firms to perform different tasks that could affect subsequent productivity. To evaluate these concerns, we measure firms' access to capital by the monetary value of the contract; we measure market potential by the GDP per capita of the prefecture where a firm's first government contract was issued; and we quantify the requirements of each contract using natural language processing,

[27] Specifically, we standardize each element of the index and combine them, weighting by the inverse covariance matrix following Anderson (2008).

[28] While firms' underlying productivity cannot account for the treatment effect we observe, one might wonder whether it is a source of heterogeneous effects of government data. To test for heterogeneous treatment effects associated with firms' underlying productivity, we estimate the baseline specification on samples of firms split above and below the median level of the productivity index. We find positive effects of similar magnitudes among both samples of firms (see Supplementary Appendix Figure A.12).

measuring the distance between the language used in each contract and a random set of non-public security contracts. We construct an index of the non-data benefits of the contract combining these characteristics (again, following Anderson, 2008), and we control for the time-varying effects of this index. Formally, we interact this index with a full set of time period fixed effects ($\sum_T T_{it} C_i$). As presented in Figure 5B, we find that these controls do not affect our estimates.

Additionally, it is possible that receipt of a data-rich contract may function as a signal of firm quality or potential through which firms could derive additional productive inputs.[29] To test whether the differential signalling value of data-rich contracts accounts for our findings, we examine the effects of a firm's first contract but limiting our analysis to subsidiary firms belonging to a mother firm that has *already* received a government contract through a different subsidiary. Arguably, the signalling value of these first contracts should be lower (mother firm quality is already observed), while access to data remains potentially extremely valuable. In Figure 5B, one can see that within this sample of firms there is still a significant differential effect of receiving a data-rich contract.

One may also be concerned that receipt of data-rich procurement contracts may be a result of firms' political connections, may strengthen such connections, and may stimulate firms' software production through these connections rather than access to government data. Connections may be differentially valuable with the local governments of Beijing and Shanghai, two specific prefectures that are highly politically significant and exhibit high levels of surveillance in most time periods. To rule out the possibility that our findings are driven by contracts with these two local governments, we estimate our baseline specification, but excluding contracts with Beijing and Shanghai governments. Our findings are qualitatively unchanged (again, see Figure 5B). Another possibility is that local firms may be able to leverage advantages with local government officials to acquire data-rich contracts and to successfully market commercial software due to stronger political ties. To rule this out, we estimate our baseline model, but excluding contracts signed between firms and any government in their home province. We again find that our results are unaffected (again, see Panel B).

4.4. *Assessing mechanisms*

As shown in the conceptual framework, government data could stimulate commercial innovation through two mechanisms: either directly, due to the shareability of data and algorithms across uses; or indirectly, through the development of non-data shareable inputs as a result of access to greater amounts of government data (e.g. improved firm capacity to manage and utilize data as a form of firm learning by doing).

We are able to measure one important dimension of non-data shareable inputs: the development of data-complementary (non-AI) software that facilitates more efficient data storage, transmission, and management. Such shareable inputs might arise from firms' learning by doing as a result of access to unprecedented quantities of government data. To examine whether this shareable input differentially responds to the receipt of data-rich public security contracts, we estimate our baseline model in Section 4.2, but considering these data-complementary software products as the outcome of interest. We present the estimates in Figure 6A. One can see that data-complementary software production differentially increases after the receipt of a data-rich public security contract. We find no evidence of pre-contract differences in data-complementary software production levels

[29] Perhaps firms obtaining data-rich government contracts receive additional benefits from local industrial policy compared to firms obtaining data-scarce ones; or attract additional external funding, human capital, or customers, all of which contribute to the production of software.

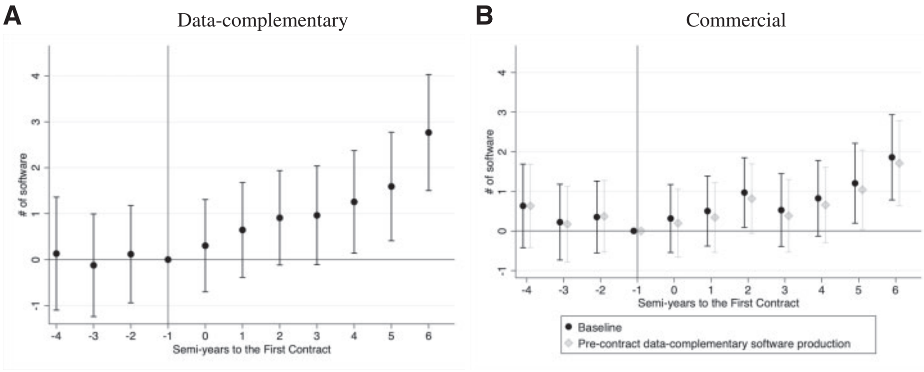

FIGURE 6

(A) replicates results from Figure 3A, but on data-complementary software releases instead. (B) replicates results from Figure 3A (plotted in dots), and also shows results (plotted in diamonds) controlling for pre-contract data-complementary software interacted with time indicators.

or trends; this suggests that receipt of a data-rich contract indeed generates shareable inputs that may induce greater commercial software production.

We next evaluate whether the indirect effect of government data through the development of non-data shareable inputs plays a predominant role in driving the observed increase in commercial AI software production. Specifically, we proxy for firms' potential to benefit from additional shareable inputs following the receipt of data-rich contracts using their *pre-contract* production of data-complementary software. To account for this potential, we estimate our baseline model (with commercial software as the outcome), additionally controlling for pre-contract data-complementary software production interacted with the full set of time period fixed effects. If potential benefits from additional non-data shareable inputs account for our baseline findings, then these controls would significantly alter our estimates. However, one can see in Figure 6B, that these controls have very little effect. This suggests an important *direct effect* of shareable government data (and algorithms) on commercial software production.

5. CONCLUSION

In this article, we provide the first evidence of a causal effect of government data on commercial AI innovation. We argue that an important mechanism underlying this effect is that data and trained algorithms are shareable across government and commercial uses. Within our empirical context, our findings suggest that the provision of government data to Chinese AI firms servicing the state contributed to their rise as global leaders in facial recognition technologies. More generally, the economic mechanism that we highlight could apply to a range of other important domains where government data are predominant—geospatial and health data being two salient examples.[30]

[30] Geospatial data collected by government satellites is used in applications related to transportation, mineral extraction, and energy production. Health data is collected by states in enormous quantities and is extremely valuable for AI-fuelled diagnoses and treatment of disease. More concretely, the British National Health Service (NHS) recently signed a contract with Amazon for AI medical services. Developing these requires Amazon to access NHS medical data which is not publicly available and could contribute to the development of Amazon's commercial AI products. See: https://bit.ly/3hNGTbT.

This implies that states' AI procurement and data provision can act as innovation policies that, intentionally or not, could shape the development of AI in many areas.[31]

Further work is needed to fully understand the normative tradeoffs involved in such policies. All states engage in surveillance to ensure public safety and security. In the modern world, this is likely to involve substantial government data collection and analysis using AI. Similarly, AI technology may be deployed to enhance the effectiveness of public health policies.[32] Adding to these direct benefits are the potential commercial AI innovation spillovers that we document. However, states' deployment of AI and data-related policies also present distinct costs. States' deployment of AI can potentially infringe on civil liberties, particularly in the case of government surveillance; government data collection and provision may run the risk of violating privacy. Evaluating the normative implications of state AI and data-related policies thus requires measuring such costs—a task complicated by the fact that they are likely to vary across societies with different values and cultural norms.

Finally, our evidence raises questions regarding political economy aspects of data-intensive innovation. Because surveillance states—particularly autocracies—collect enormous amounts of data to monitor their citizens, one naturally wonders whether they may exhibit rapid AI innovation despite their repressive and extractive institutions. At the extreme, might surveillance states and societies with weaker privacy norms have a comparative advantage in AI innovation, and if so, what are the implications for trade policy? Answers to these questions will help us understand the consequences of China's rise as an AI superpower, and more generally, the global economic *and* political landscape in the age of data-intensive innovation.

Acknowledgments. We are especially grateful for the extraordinary research assistance provided by Haoran Gao, Andrew Kao, Shuhao Lu, and Wenwei Peng. We also thank Shiyun Hu, Junxi Liu, Shengqi Ni, Yucheng Quan, Linchuan Xu, Peilin Yang, and Guoli Yin, for their excellent work as research assistants as well. Many appreciated suggestions, critiques and encouragement were provided by Daron Acemoglu, Dominick Bartelme, Ryan Bubb, Paco Buera, Ernesto Dal Bó, Dave Donaldson, Ruben Enikolopov, Raquel Fernández, Richard Freeman, Chad Jones, Pete Klenow, Monica Martinez-Bravo, Andy Neumeyer, Juan Pablo Nicolini, Arianna Ornaghi, Maria Petrova, Torsten Persson, Nancy Qian, Andrei Shleifer, Chris Tonetti, Dan Trefler, John Van Reenen, and Daniel Xu, as well as many seminar and conference participants. Yuchtman acknowledges financial support from the British Academy under the Global Professorships program.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online. And the replication packages are available at https://dx.doi.org/10.5281/zenodo.6607692.

Data Availability Statement

The data and code underlying this research is available on Zenodo at https://dx.doi.org/10.5281/zenodo.6607692.

REFERENCES

ACEMOGLU, D. and RESTREPO, P. (2019), "The Wrong Kind of AI? Artificial Intelligence and the Future of Labour Demand", *Cambridge Journal of Regions, Economy and Society*, **13**, 25–35.

AGHION, P., JONES, B. F. and JONES, C. I. (2019), *9. Artificial Intelligence and Economic Growth* (Chicago, IL, USA: University of Chicago Press).

AGRAWAL, A., GANS, J. and GOLDFARB, A. (2018), *Prediction Machines The Simple Economics of Artificial Intelligence* (Boston, MA, USA: Harvard Business Press).

——————, ——————, and ——————, eds, (2019), *The Economics of Artificial Intelligence An Agenda* (Chicago, IL, USA: University of Chicago Press).

[31] Note, however, that government provision of data to specific firms could distort the competitive landscape and discourage entry, thus dampening the overall growth of the sector.

[32] The US CDC writes that AI technology, "could support public health surveillance, research and, ultimately, decision making". See https://bit.ly/2WcfTKD.

ALIC, J. A., BRANSCOMB, L. M., BROOKS, H. and CARTER, A. B. (1992), *Beyond Spinoff Military and Commercial Technologies in a Changing World* (Boston, MA, USA: Harvard Business Press).

ANDERSON, M. L. (2008), “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects”, *Journal of the American Statistical Association*, **103**, 1481–1495.

ARIDOR, G., CHE, Y.-K., NELSON, W. and SALZ, T. (2020), “The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR” (Working Paper) 1–67.

AZOULAY, P., FUCHS, E., GOLDSTEIN, A. P. and KEARNEY, M. (2019), “Funding Breakthrough Research: Promises and Challenges of the “ARPA Model”,” *Innovation Policy and the Economy*, **19**, 69–96.

BAI, C.-E., HSIEH, C.-T. and SONG, Z. (2020), “Special Deals with Chinese Characteristics”, *NBER Macroeconomics Annual*, **34**, 341–379.

BLOOM, N., JONES, C. I., VAN REENEN, J. and WEBB, M. (2020), “Are Ideas Getting Harder to Find?”, *American Economic Review*, **110**, 1104–1144.

———, VAN REENEN, J. and WILLIAMS, H. L. (2019), “A Toolkit of Policies to Promote Innovation”, *Journal of Economic Perspectives*, **33**, 163–184.

BOMBARDINI, M., LI, B. and WANG, R. (2018), “Import Competition and Innovation: Evidence from China” (Working Paper) 1–44.

BORUSYAK, K., JARAVEL, X. and SPIESS, J. (2017), “Revisiting Event Study Designs: Robust and Efficient Estimation” (Working Paper) 1–35.

BRANDT, L. and RAWSKI, T. G. (2008), *China’s Great Economic Transformation* (Cambridge, UK: Cambridge University Press).

CHEN, Y., EBENSTEIN, A., GREENSTONE, M. and LI, H. (2013), “Evidence on the Impact of Sustained Exposure to Air Pollution on Life Expectancy from China’s Huai River Policy”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 110. 12936–12941.

CHENG, H., JIA, R., LI, D. and LI, H. (2019), “The Rise of Robots in China”, *Journal of Economic Perspectives*, **33**, 71–88.

CLEMENS, J. and ROGERS, P. (2020), “Demand Shocks, Procurement Policies, and the Nature of Medical Innovation: Evidence from Wartime Prosthetic Device Patents” (NBER Working Paper) 1–94.

FARBOODI, M., MIHET, R., PHILIPPON, T. and VELDKAMP, L. (2019), “Big data and firm dynamics,” in Johnson, W. (ed) *AEA Papers and Proceedings*, vol. 109. 38–42. https://www.aeaweb.org/journals/pandp/about-pandp.

FERRAZ, C., FINAN, F. and SZERMAN, D. (2015), “Procuring Firm Growth: The Effects of Government Purchases on Firm Dynamics” (NBER Working Paper 21219).

FISMAN, R. and WANG, Y. (2015), “The Mortality Cost of Political Connections”, *The Review of Economic Studies*, **82**, 1346–1382.

GREENSTEIN, S. (2015), *How the Internet Became Commercial Innovation, Privatization, and the Birth of a New Network* (Princeton, NJ, USA: Princeton University Press).

GROSS, D. P. and SAMPAT, B. N. (2020), “Inventing the Endless Frontier: The Effects of the World War II Research Effort on Post-War Innovation” (NBER Working Paper) 1–58.

HE, G., WANG, S. and ZHANG, B. (2020), “Watering Down Environmental Regulation in China”, *The Quarterly Journal of Economics*, **135**, 2135–2185.

JIA, R., KUDAMATSU, M. and SEIM, D. (2015), “Political Selection in China: the Complementary Roles of Connections and Performance”, *Journal of the European Economic Association*, **13**, 631–668.

JONES, C. I. and TONETTI, C. (2020), “Nonrivalry and the Economics of Data”, *American Economic Review*, **110**, 2819–2858.

KHANDELWAL, A. K., SCHOTT, P. K. and WEI, S.-J. (2013), “Trade Liberalization and Embedded Institutional Reform: Evidence from Chinese Exporters”, *American Economic Review*, **103**, 2169–2195.

LAU, L. J., QIAN, Y. and ROLAND, G. (2000), “Reform without Losers: An Interpretation of China’s Dual-Track Approach to Transition”, *Journal of Political Economy*, **108**, 120–143.

LI, H. and ZHOU, L.-A. (2005), “Political Turnover and Economic Performance: The Incentive Role of Personnel Control in China”, *Journal of Public Economics*, **89**, 1743–1762.

LI, W. (2019), “Rotation, Performance Rewards, and Property Rights” (Working Paper) 1–75.

MORETTI, E., STEINWENDER, C. and VAN REENEN, J. (2019), “The Intellectual Spoils of War? Defense R&D, Productivity and International Spillovers” (NBER Working Paper) 1–76.

NAGARAJ, A. (2021), “The Private Impact of Public Data: Landsat Satellite Maps Increased Gold Discoveries and Encouraged Entry”, *Management Science*, **68**, 564–582.

NAGLE, F. (2019), “Government Technology Policy, Social Value, and National Competitiveness” (Working Paper) 1–52.

PERRAULT, R., SHOHAM, Y., BRYNJOLFSSON, E., CLARK, J., ETCHEMENDY, J., GROSZ, B., LYONS, T., MANYIKA, J., MISHRA, S. and NIEBLES, J. C. (2019), “The AI Index 2019 Annual Report” (Technical Report, AI Index Steering Committee, Human-Centered AI Institute, Stanford University).

ROBERTS, M. J., YI XU, D., FAN, X. and ZHANG, S. (2017), “The Role of Firm Factors in Demand, Cost, and Export Market Selection for Chinese Footwear Producers”, *The Review of Economic Studies*, **85**, 2429–2461.

SCOTT, J. C. (1998), *Seeing Like a State How Certain Schemes to Improve the Human Condition Have Failed* (New Haven, CT, USA: Yale University Press).

SLAVTCHEV, V. and WIEDERHOLD, S. (2016), “Does the Technological Content of Government Demand Matter for Private R&D? Evidence from US States”, *American Economic Journal: Macroeconomics*, **8**, 45–84.

SONG, Z., STORESLETTEN, K. and ZILIBOTTI, F. (2011), “Growing Like China”, *American Economic Review*, **101**, 196–233.

TIROLE, J. (2021), “Digital Dystopia”, *American Economic Review*, **111**, 2007–2048.

WEI, S.-J., XIE, Z. and ZHANG, X. (2017), “From “Made in China” to “Innovated in China”: Necessity, Prospect, and Challenges”, *Journal of Economic Perspectives*, **31**, 49–70.

WILLIAMS, H. L. (2013), “Intellectual Property Rights and Innovation: Evidence from the Human Genome”, *Journal of Political Economy*, **121**, 1–27.