NBER WORKING PAPER SERIES

DATA-INTENSIVE INNOVATION AND THE STATE: EVIDENCE FROM AI FIRMS IN CHINA

Martin Beraja
David Y. Yang
Noam Yuchtman

Working Paper 27723
http://www.nber.org/papers/w27723

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2020, Revised March 2021

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 Free-man, 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. Yang acknowledges financial support from the Harvard Data Science Initiative; Yuchtman acknowledges financial support from the British Academy under the Global Professorships program. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Data-intensive Innovation and the State: Evidence from AI Firms in China
Martin Beraja, David Y. Yang, and Noam Yuchtman
NBER Working Paper No. 27723
August 2020, Revised March 2021
JEL No. E0,H4,L5,L63,O25,O30,O40,P00,P16,Z21

ABSTRACT

Artificial intelligence (AI) innovation is data-intensive. States have historically collected large amounts of data, which is now being used by AI firms. Gathering comprehensive information on firms and government procurement contracts in China's facial recognition AI industry, we first study how government data shapes AI innovation. We find evidence of a precise mechanism: because data is sharable across uses, economies of scope arise. Firms awarded public security AI contracts providing access to more government data produce more software for both government and commercial purposes. In a directed technical change model incorporating this mechanism, we then study the trade-offs presented by states' AI procurement and data provision policies. Surveillance states' demand for AI may incidentally promote growth, but distort innovation, crowd-out resources, and infringe on civil liberties. Government data provision may be justified when economies of scope are strong and citizens' privacy concerns are limited.

Martin Beraja
Department of Economics, E52-504
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
martinberaja@gmail.com

Noam Yuchtman
London School of Economics
Houghton St.
London WC2A 2AE
United Kingdom
and CEPR
n.yuchtman@lse.ac.uk

David Y. Yang
Department of Economics
Harvard University
Littauer Center M-31
Cambridge, MA 02138
and NBER
davidyang@fas.harvard.edu

1 Introduction

Developing artificial intelligence and machine learning technologies ("AI" for brevity) is *data-intensive*. Up to now, economists have emphasized how data collected by private firms shapes the process of AI innovation (Agrawal et al., eds, 2019; Jones and Tonetti, 2018). Yet, throughout history and up to the present, states have also collected massive quantities of data in order to fulfill their objectives of revenue extraction and public goods provision (Scott, 1998). Private firms providing goods and services to the state are often able to access government data, and such data is currently used to train algorithms in many leading AI applications: administrative health records are used for medical diagnoses; geospatial data are used to detect mineral resources; satellite and radar data are used in meteorological prediction; and video from public surveillance cameras is used in facial recognition.

In this paper, we ask: how does government data shape AI innovation? And, what are the trade-offs presented by states' AI procurement and data provision policies in the age of data-intensive innovation? We first highlight a precise mechanism linking government data to private AI innovation: because data can be shared across multiple uses within a firm (Goldfarb and Trefler, 2018), firms gaining access to government data after obtaining a contract to supply AI software to the state could use that same data to develop not only products for government uses but also products intended for much larger commercial markets.[1] We document the existence of such *economies of scope* in the context of a leading AI sector in a country at the technological frontier: the facial recognition AI sector in China.[2] Within the set of AI contracts with public security agencies, we measure government data access using the number of cameras in the agency's surveillance network that can capture high-resolution video of faces on the street. We find that obtaining a contract from an agency with more cameras in its local surveillance network causes firms to produce more AI software for *both* government and commercial purposes. To the best of our knowledge, this is the first causal evidence of the effect of government data on firms' AI innovation.[3]

[1] Panzar and Willig (1981) show how economies scope may arise when inputs are sharable. The sharability of data across multiple uses within the firm is related to the non-rivalry of data across firms, which has been highlighted by Jones and Tonetti (2018), among others.

[2] Facial recognition AI is among the top three AI technologies in terms of projected revenues (Perrault et al., 2019). China is the world's largest producer of AI research (see the "China AI Development Report, 2018," available online at `https://bit.ly/2IWAo7R`.

[3] This evidence also contributes to the literature on how the Chinese state affects economic activity (e.g., Lau et al., 2000; Brandt and Rawski, 2008; Song et al., 2011), and specifically, how autocracy can foster economic growth (e.g., Bai et al., 2019).

Then, we analyze the trade-offs presented by two state policies closely linked to our empirical context and at the center of recent policy discussions. To do so, we build a directed technical change model with data as an input and economies of scope generated by government data. The first policy concerns states' demand for AI technologies to monitor their citizens and provide public security — with extreme manifestations being surveillance states' or informational autocrats' (Guriev and Treisman, 2019) demand for AI. We show that demand for AI to support a surveillance state may *incidentally* promote economic growth, but distort the direction of innovation and reduce citizen welfare due to the crowding-out of resources from consumption and the infringement of civil liberties. The second policy concerns states' direct provision of government data to firms, similar to industrial and innovation policies subsidizing inputs in order to promote certain sectors and technologies (Rodrik, 2007; Bloom et al., 2019). We show that government data provision to firms may be justified as a form of innovation policy when economies of scope are strong and citizens' privacy concerns about the collection and sharing of their data are limited. Taken together, these results suggest that states' demand for AI and data provision to firms may stimulate AI innovation just as their spending on space exploration and national defense stimulated innovation in the past (Azoulay et al., 2018a; Moretti et al., 2019; Gross and Sampat, 2020), though working through distinct mechanisms and implying distinct trade-offs.

Our paper begins by presenting a simple conceptual framework where economies of scope in data-intensive innovation can arise from government data being sharable across multiple uses — or, as we discuss, a base AI algorithm trained with such data being *transferable* across uses. AI firms receive contracts to produce government software using government data; they can use the same government data (or a transferable algorithm trained with such data) to produce commercial software as well. Yet, economies of scope may not arise even when there is sharability across uses, if government software production requires the reallocation of substantial non-data resources away from commercial software production. We thus derive a test for economies of scope that guides our subsequent empirical analysis: whether receipt of a government contract providing access to more government data leads to increased production of *both* government and commercial software.

The facial recognition AI industry in China is a particularly well suited empirical context to examine economies of scope arising from government data. Firms developing facial recognition software require large training datasets: for example, training an algorithm to match the same face observed at different angles across different video streams

requires enormous amounts of video training data.[4] The decentralized public security units of the Chinese state (e.g., municipal police departments) collect huge amounts of *precisely* this form of data through their surveillance apparatus, which is analyzed by private facial recognition AI software firms receiving contracts from the respective government units. A contracted facial recognition AI firm thus receives access to government data which is not publicly available, using this data to train AI software to satisfy the government's surveillance demand. Crucially, the matching and detection of individuals from video data is key to *both* government and commercial facial recognition AI applications (for instance, facial recognition platforms for retail stores). Therefore, to the extent that the government data (or fine-tuned, transferable algorithm) is sharable across uses, there may exist economies of scope.

Reflecting this discussion, 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 measurement challenges. First, linking AI firms to government contracts. To do so, we collect information 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. We measure the data provided by a public security contract using the agency's local surveillance network's capacity to record 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 the causal effect of access to gov-

[4]Depending on the application, firms can train algorithms using *identifiable* data (e.g., video surveillance feeds not linked to administrative records), *identified* data (e.g., linked faces and names in ID databases), or both in combination.

ernment data by comparing the cumulative increase in software releases of firms that receive data-rich and data-scarce public security contracts, respectively. By exploiting variation in data-richness within the set of public security contracts, this comparison 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. We find that receipt of a data-rich contract *differentially* increases *both* government and commercial software production, relative to receipt of a data-scarce contract. In the three years after the receipt of a contract, data-rich contracts generate an *additional* 3 government software products (over and above the effects of a data-scarce contract), and an additional 2 commercial software products. Our evidence thus indicates the presence of economies of scope, reflecting crowding-*in* rather than crowding-out.

We provide a range of corroborating evidence for our proposed mechanism of access to government data contributing to product innovation. First, 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. Second, we observe lower bids for data-rich contracts, as well as more bidders overall. Finally, we find that firms receiving data-rich public security contracts differentially produce *video* facial recognition AI software, which is particularly data-intensive.

We conclude our empirical analysis by evaluating a range of threats to identification and alternative mechanisms. First, one may be concerned of non-random assignment of contracts to firms. By including firm fixed effects and comparing the differential effects between the receipts of data-rich and data-scarce contracts, our baseline empirical strategy allows us to account for both time-invariant and time-varying factors that drive firms' selection into receipt of any public security contracts, as well as time-invariant selection into receipt of a data-rich public security contract. One might still be concerned about time-varying sources of firm selection into data-rich contracts. However, our event-study estimates show no differential software production prior to receipt of a data-rich contract, and our findings are robust to allowing pre-contract firm characteristics to flexibly affect post-contract output. Second, we provide evidence showing that our main results are also unlikely to be explained by differences between data-rich and data-scarce contracts along dimensions other than data: their terms and tasks required, potential for learning-by-doing, access to capital, signaling value, associated commercial opportunities, or connections to local government.

Having established the effect of government data on innovation at the micro (i.e., firm) level, we then study the macro implications of government data and the trade-offs pre-

sented by different state policies. We present a directed technical change model, building on Acemoglu (2002). We let innovator firms develop and supply differentiated varieties of data-intensive government and commercial software, as well as other, non-software varieties which do not use data as an input. Commercial software and non-software are used to produce a final good. Government software is purchased by the state to produce a government good, which we call "surveillance" for concreteness. A representative household owns all firms and consumes the final good. There are two types of data in the economy: government and private. Government data is necessary for producing government software. We assume that the same government data could simultaneously be used for producing both government and commercial software, generating economies of scope. Government data is produced as a by-product of surveillance, whereas private data is a by-product of total private transactions (as measured by final good output). Both types of data are excludable, but only private data can be purchased in the market. As in our empirical context, government data can only be accessed by producing government software for the state after receiving a government contract.

Given a state policy determining government spending and government data provided to firms, we show conditions under which there is a unique balanced growth path (BGP) equilibrium with free-entry of all types of innovators. We conclude by presenting comparative static exercises with respect to changes in government spending on data-intensive technologies and government data provision, which are two dimensions of state policy in our empirical context. When commercial software and non-software are sufficiently substitutable, a state's increased demand of data-intensive technologies or provision of government data to firms can bias the direction of private innovation and increase the BGP rate of economic growth. This result thus shows that our firm-level findings may carry over to the aggregate. However, the normative consequences of these policies are more ambiguous because both economic and non-economic forces may offset the welfare gains from a higher economic growth rate: *(i)* the crowding out of resources from consumption, and *(ii)* citizens' disutility from the state's use of data-intensive technologies (e.g., due to civil liberties infringement from excessive surveillance) or government data collection and sharing (e.g., due to privacy violations). Through several numerical exercises, we illustrate these trade-offs as well as how they are affected by the strength of economies of scope.

2 Related literature

Our work most directly contributes to an emerging literature on the economics of AI and data, particularly work that aims to understand the role of AI technology and data in fostering innovation, and firm and aggregate growth (see, e.g., Aghion et al., 2017; Agrawal et al., 2018; Farboodi et al., 2019; Acemoglu and Restrepo, 2019). We contribute to this literature by examining the role of *government*-collected data and the direct and indirect ways in which data-intensive innovation may be shaped by the state. Our analysis complements a recent literature studying the effects of specific characteristics of information and data on innovation. Aghion et al. (2017) and Jones and Tonetti (2018) study non-rivalry of data across firms. Williams (2013) and Nagaraj and Stern (2020) study settings in which the non-excludability of government research — mapping the genome and mapping the Earth, respectively — shape private sector outcomes. We instead emphasize the economies of scope arising from the sharability of government data across government and commercial applications within a firm.

A second closely-related literature studies the indirect innovation consequences of government efforts to develop technology, from space exploration (Alic et al., 1992; Azoulay et al., 2018a), to military technology (Greenstein, 2015; Moretti et al., 2019; Gross and Sampat, 2020). Our proposed mechanism shares much with this work: government data (or trained algorithms), like scientific ideas and other intangible assets, can be shared across uses and generate economies of scope. Moreover, government AI procurement and data provision to firms are also not necessarily pursued with the primary aim of promoting commercial innovation, but perhaps other political and strategic objectives. Our work builds on this literature, and we highlight the characteristics that government data shares with the basic science and technology developed by states (especially by their military units). Empirically, we identify a specific, causal mechanism through which a sharable input affects commercial innovation at the firm level. Finally, we theoretically explore the trade-offs presented by states' AI procurement and data provision policies in a world in which data-intensive innovation is increasingly important.

Our examination of the link between the state and the private sector AI industry also contributes to literatures on both industrial policy and innovation policy. Rodrik (2007) and Lane (2020) provide recent overviews of the industrial policy literature, with the latter highlighting quasi-experimental evidence of effective industrial policy.[5] Recent re-

[5]Contexts in which industrial policy was shown to be effective include: the 19th century French textile industry, protected by the blockade of British competitors during the Napoleonic Wars (Juhàsz, 2018); 19th century UK and Great Lakes US shipbuilding (Hanlon, 2020); post-WWII Finland following industrialization imposed by the Soviet Union (Mitrunen, 2019); post-WWII Italy, as a result of the US Marshall Plan

search on innovation policy also suggests an important role for the state in encouraging R&D — see Bloom et al. (2019).[6] We make three primary contributions to these literatures. First, we study a frontier technology: the effects of the state on the development of modern AI innovation, a technology which has enormous economic potential, and which also may be particularly sensitive to state policy. Second, we conceptualize and empirically identify a specific within-firm mechanism underlying spillovers from government expenditure to private innovation in our setting. We highlight that economies of scope across government and commercial uses could generate consequences similar to those achieved by industrial policy and innovation policy, despite the incidental nature of the state's engagement, for example, due to states' demand for surveillance or due to citizens' demand for privacy protection.[7] Third, we provide a justification for government data provision that differs from that of traditional industrial policies. For example, Costinot et al. (2019) evaluate the case for industrial policy to correct for learning-by-doing externalities. We show that because states are key collectors of data, and because government data can give rise to economies of scope, it may be optimal to directly provide such data to data-intensive software producers, even in the absence of externalities. In this sense, we also contribute to a macroeconomic literature on the role of government spending in promoting economic growth (e.g., Murphy et al., 1989, Barro, 1990).

By placing our analysis of AI innovation within a model of directed technical change, we contribute to the body of work on these models (e.g., Acemoglu, 1998; Acemoglu et al., 2012; Hemous, 2016). We add to this literature by studying a novel application — data-intensive innovation and the role of the state. Our empirical analysis contributes to a much smaller body of empirical work on directed technical change (Popp, 2002; Acemoglu et al., 2006; Hanlon, 2015; Aghion et al., 2016; Costinot et al., 2019). We add to this literature by documenting how an increase in the supply of data, as a result of receiving a government contract, induces Chinese firms to develop (data-intensive) commercial applications of AI technologies.

Finally, we highlight the political dimension of data-intensive AI innovation. Data is valued — and thus accumulated — by modern surveillance states, particularly by auto-

(Giorcelli, 2019); East Asia's (and China's) growth miracle (Lane, 2017; Liu, 2019); and, Chinese shipbuilding in the 2000s (Kalouptsidi, 2017; Barwick et al., 2019). Bartelme et al. (2019) estimate the importance of sectoral economies of scale that are often used to justify industrial policy, finding that industrial policy may not be as effective as other policies (e.g., trade).

[6] Among others, Howell (2017) shows that the US Department of Energy's funding helps firms to innovate; Azoulay et al. (2018b) show that public grants increase patenting by pharmaceutical and biotechnology firms; and Moser (2005) studies how intellectual property rights shape innovation.

[7] 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).

cratic states (Tirole, 2020). In addition, a fundamental aim of AI technology — to make accurate predictions — is aligned with their surveillance and social control agenda (Guriev and Treisman, 2019; Zuboff, 2019). Therefore, AI is a technology that can buttress rather than threaten autocratic regimes. Combining these insights, our project contributes to our understanding of how political economy affects the rate and direction of technical change. Traditionally, scholars have emphasized limits on entrepreneurship under autocracies arising from the misaligned incentives facing entrepreneurs and political elites.[8] In the domain of AI technology, however, surveillance states' objectives and data collection, along with the economies of scope arising from data as an input, facilitate data-intensive innovation even for commercial applications. Thus, the alignment between the state and private sector could offset the expropriation risks and commitment problems traditionally faced by private entrepreneurs under autocracy, although, as we emphasize, such alignment may still be detrimental to citizens overall. Our analysis thus may also help explain the puzzle of China's global leadership in AI innovation and more generally suggests that modern autocracy may be compatible with technical change along specific trajectories. In so doing, we contribute to a nascent literature (e.g., Bai et al., 2019) that identifies a mechanism through which autocratic power can actually *promote* economic growth.[9]

3 Economies of scope from government data

Suppose that a firm may develop data-intensive software for both the state and the private sector. Assume that developing software for the state uses government data d_g as an input. Imagine — as is the case in reality — that there exist types of government data that lack close substitutes (e.g., surveillance video from street cameras) and that are not made publicly available.[10] In order to obtain access to these types of government data, the firm must obtain a contract from the state to produce government software. Government

[8]The risk of *ex post* taxation or expropriation of entrepreneurs induces *ex ante* less investment (North et al., 2009; Acemoglu and Robinson, 2012). Threats to elites arising from successful entrepreneurs also lead elites to *ex ante* tax entrepreneurs to preserve their political rents (Acemoglu and Robinson, 2006). Public sector distortions such as corruption may also discourage innovation and investment (Shleifer and Vishny, 2002).

[9]A large literature studies the Chinese economy and its spectacular growth in the recent decades (e.g., Song et al., 2011; Khandelwal et al., 2013; Roberts et al., 2017; Cheng et al., 2019), as well as innovation in China more specifically (e.g., Wei et al., 2017; Bombardini et al., 2018). Much of the work on China's political economy highlights various distortions caused by the state (e.g., Chen et al., 2013; Fisman and Wang, 2015; He et al., 2020), and institutional features that allow China to grow despite the lack of institutional constraints on the Chinese Communist Party — for example, competition for promotion (e.g., Li and Zhou, 2005; Jia et al., 2015), bureaucratic rules of evaluation and rotation (Li, 2019), or social norms (Tsai, 2007).

[10]Other examples of potentially valuable government data include personally-identified health records and data on earnings, as well as geographic and geological data, among others.

software production also uses a number of other inputs, including other forms of data, which can be purchased in the market, and which we denote in vector form by x_g. Then, we let $F_g(d_g, x_g)$ be the production function of government software S_g.

Moreover, assume that if a firm has access to government data d_g, then it can use that *same* data to produce commercial software for the private sector. That is, government data can be *shared across uses*. We let $F_c(d_g, x_c)$ be the production function of commercial software S_c, where x_c is again a vector of other types of inputs. As an example of government data and its shared uses, consider video from street surveillance cameras and administrative records with the names of individuals linked to images of their faces. This data is used to train an algorithm with the ability to *recognize* faces in video and identify individuals in administrative records. That trained identification algorithm may then also be part of a more complex software application that performs the *predictive* task of identifying potential security threats. That same data, though, is also a crucial input to train algorithms that perform a wide range of *commercial* recognition and prediction tasks, such as identifying a customer in video from store cameras or predicting their purchases.

An alternative plausible specification of the technologies is one where government data is not shared across uses *per se*, but the data is instead used to train a "base algorithm" which is *transferable* and can thus itself be used as an input to develop both government and commercial software.[11] We will treat the sharability of data or trained algorithms with it as equivalent, because, for the purposes of this paper, it is immaterial whether the value of government data for commercial innovation is derived from the sharability of the data, or from transferable algorithm better trained with such data.

Following Panzar and Willig (1981), it is possible that *economies of scope* arise when $\frac{\partial F_c}{\partial d_g} > 0$. Intuitively, this is because the firm obtaining more government data by producing government software could produce a given level of commercial software S_c with less of the other inputs, and thus at lower cost.[12] This generates a testable implication about the firm-level impact of obtaining a government contract that is richer in data, when there are economies of scope. Consider a firm that is already producing commercial software. Suppose it receives a government contract to produce government software, which provides access to government data (with $\frac{\partial F_c}{\partial d_g} > 0$). Then this firm could begin to produce not only more government software (using government data), but also more commer-

[11] Indeed, in machine learning, the subfields of transfer learning and domain adaptation are specifically devoted to studying problems related to sharability of trained algorithms and data across uses.

[12] Imagine that the firm splits in two: one only producing government software (with access to government data) and the other one only producing private software (without access to government data). Formally, let input prices be ω and let $C(S_g, S_c, d_g, \omega)$, $C_g(S_g, 0, d_g, \omega)$, and $C_c(0, S_c, 0, \omega)$ be the cost functions of the firms producing both types of software and each type separately. Then, there are economies of scope when $C(S_g, S_c, d_g, \omega) < C_g(S_g, 0, d_g, \omega) + C_c(0, S_c, 0, \omega)$.

cial software, because the government data to which it receives access can be used for commercial software production as well.

Note, however, that these economies of scope are not guaranteed. For instance, when a firm uses resources to produce more government software, this may *crowd-out* resources that would have been used for commercial software production. If such crowding-out effects are relatively strong, obtaining a government contract that is richer in government data would induce the firm to produce more government software but *less* commercial software. Observing increases in *both* government and commercial software production following receipt of a data-rich government contract would thus be strong evidence for economies of scope arising from government data, where the ability to share data (or the algorithm trained with it) across uses more than offsets any crowding out of resources.

In the next section, we test for this implication of economies of scope in the context of China's AI industry:

Implication of economies of scope arising from government data: Obtaining a government contract that is richer in government data induces a firm to produce both more government and commercial software.

4 The state and China's facial recognition AI industry

4.1 Empirical context

China's facial recognition AI sector is particularly well suited to examine the impact of access to government data on innovation and to provide evidence of economies of scope arising from such data. First, because facial recognition AI is extremely data-intensive: the development of the technology requires access to large datasets containing faces. Second, because the Chinese state collects huge amounts of surveillance data and demands AI software in order to monitor citizens. 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 is the biggest source of data for AI firms like us."[13] 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."[14]

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

[14]*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

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 includes video from street surveillance cameras, and, potentially, labeled images with names and faces of individuals. The firm uses this data to train an AI algorithm; e.g., a "tracking" algorithm that matches faces across video feeds or a "detection" algorithm that matches faces from video to the database of individuals. Then, economies of scope can arise from the government data (or a base algorithm trained with it) being 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.

This context allows us to empirically test for economies of scope arising from access to government data. In particular, in the next section we exploit within-firm variation over time in the receipt of procurement contracts, together with variation in the data available to firms under different contracts. This allows us to estimate the effect of access to more government data on both government and commercial software production.

4.2 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 firms' software production, and specifically whether a given software is intended for government or commercial 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.[15]

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 licensed by China's central bank.[16] 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

broadly a characteristic of AI innovation around the world.

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

[16] See Appendix Figure A.1 for an example entry.

from *Pitchbook*, a database owned by Morningstar on firms and private capital markets around the world.[17] 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 7,837 Chinese facial recognition AI firms.[18] 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.[19] 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. As an example from our dataset, consider a 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 August 29th, 2018. The contract specifies that the AI firm shall provide a facial recognition system that should cover at least *30 million* individuals, suggesting the large scale of data collection and processing that are required.

We begin with a comprehensive set of public security agency procurement contracts, including 410,510 contracts in total. 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.

Within this set, 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.[20] Many firms receive

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

[18] These firms fall into 3 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).

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

[20] We present the cumulative number of AI procurement contracts in 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

multiple contracts; overall, 1,095 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.[21] 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*.[22] Reflecting the economic value of such innovation, we observe that facial recognition AI firms that develop more software have significantly and substantially higher market capitalization (see 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 5.2.3).

We use a Recurrent Neural Network (RNN) model with tensorflow — a frontier method for analyzing 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

security AI contracts have steadily increased since 2013.

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

[22]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 et al. (2020).

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 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. Appendix Figure A.6 shows the summary statistics of the categorization output by customers and by function; and, Appendix Figure A.7 presents the confusion matrix (Type-I and Type-II errors) of the predictions relative to categorization done by humans.[23]

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.[24] Specifically, using 5,837 prefectural government contracts for purchases of surveillance cameras, we

[23]Appendix Table A.2 presents the top words (in terms of frequency) used for the categorization. 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.

[24]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.

Figure 1: Circle size indicates the number of first public security 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.

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.[25]

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 1 shows the distribution of data-rich and data-scarce contracts across prefectures according to our definition.[26] 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 differences (both observable and unobservable) between the firms receiving

[25]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. This is affirmed in the Chinese central government's official directive on public security video surveillance; see: `https://bit.ly/3dqdjU0`. There are on average 77 surveillance camera contracts per prefecture. In Appendix Figure A.9, we present a time series plot of the number of cameras in our data over time. We normalize the camera counts by local population size to capture the idea that multiple observations per individual are particularly valuable for improving facial recognition AI accuracy; the results we present are robust to using total camera counts instead.

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

Table 1: Summary statistics — firms and their production

	Received at least one government contract		Received at least one public security contract		Data-richness of Public security contract	
	Yes	No	Yes	No	High	Low
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Firm characteristics						
Year firm established	2,009.335	2,013.781	2,008.947	2,011.385	2,007.506	2,009.994
	(6.389)	(4.244)	(6.376)	(6.071)	(6.963)	(5.696)
Capitalization (millions USD)	22.964	5.091	26.552	4.136	35.366	20.096
	(210.840)	(43.007)	(229.816)	(14.364)	(295.412)	(166.131)
Rounds of investment funding	2.517	2.046	2.599	1.714	2.746	2.500
	(1.961)	(3.258)	(2.013)	(1.073)	(2.075)	(1.969)
Observations	1,093	6,041	919	174	387	532
Panel B: Software production before contract						
Total amount of software	22.653	14.572	23.787	14.826	27.441	21.214
	(37.860)	(24.473)	(39.905)	(16.409)	(44.955)	(35.752)
Commercial	9.020	6.283	9.352	6.727	10.096	8.829
	(17.087)	(12.502)	(17.894)	(9.602)	(20.135)	(16.131)
Government	7.342	3.951	7.807	4.132	9.959	6.292
	(16.269)	(8.175)	(17.156)	(6.989)	(17.678)	(16.630)
AI-common	3.878	2.580	4.056	2.645	4.397	3.816
	(7.300)	(6.280)	(7.650)	(3.951)	(8.082)	(7.329)
AI-video	1.564	1.026	1.587	1.405	2.003	1.294
	(3.838)	(2.827)	(3.918)	(3.242)	(4.894)	(3.021)
Data-complementary	9.235	5.632	9.726	5.851	11.255	8.649
	(16.704)	(10.763)	(17.535)	(8.383)	(19.423)	(16.007)
Observations	956	6,042	835	121	345	490
Panel C: Software production after contract						
Total amount of software	24.393		28.239	4.171	37.395	21.591
	(59.812)	-	(64.395)	(10.101)	(80.274)	(48.810)
Commercial	8.381		9.662	1.646	12.336	7.720
	(19.628)	-	(21.101)	(4.032)	(27.498)	(14.545)
Government	9.105		10.580	1.349	14.429	7.786
	(34.029)	-	(36.909)	(3.672)	(43.605)	(30.904)
AI-common	4.126		4.804	0.560	6.256	3.750
	(11.623)	-	(12.538)	(1.973)	(16.581)	(8.333)
AI-video	1.584		1.828	0.303	2.362	1.441
	(4.014)	-	(4.310)	(1.117)	(5.057)	(3.630)
Data-complementary	10.046		11.592	1.914	15.499	8.756
	(26.230)	-	(28.266)	(5.185)	(36.313)	(20.105)
Observations	1,095	0	920	175	387	533

Note: Observations at the firm level. Standard deviations are reported below the means.

data-rich and data-scarce contracts will be crucial to identify the effects of the contracts. Table 2 presents summary statistics describing the contracts procuring AI in our sample.[27] 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 also affect software production will be crucial to identifying the causal effects of interest.

[27] In Appendix Table A.3, we provide descriptive statistics for the prefectures where contracts were issued, again disaggregating by the type of agency and by surveillance capacity.

Table 2: Summary statistics — procurement contracts

	Non-public security contracts	Public security contracts		
	All	All	Data-scarce	Data-rich
	(1)	(2)	(3)	(4)
Panel A: All contracts				
Admin level: provincial or above	0.340	0.277	0.138	0.306
	(0.474)	(0.448)	(0.345)	(0.461)
Year contract signed	2,016.350	2,016.199	2,016.274	2,016.360
	(1.612)	(1.604)	(1.516)	(1.530)
Area GDP	4,248.551	3,931.975	2,629.278	5,379.756
	(4,979.406)	(4,567.528)	(3,364.656)	(5,272.500)
Area population	479.825	480.804	404.782	569.690
	(264.595)	(263.863)	(221.149)	(284.979)
Cameras per million residents	4.311	3.392	0.138	6.920
	(8.914)	(7.493)	(0.321)	(9.644)
Observations	15,523	10,677	4,880	4,500
Panel B: First contracts				
Admin level: provincial or above	0.462	0.383	0.272	0.423
	(0.499)	(0.487)	(0.447)	(0.496)
Year contract signed	2,015.935	2,015.594	2,015.893	2,015.920
	(1.840)	(1.976)	(1.883)	(1.875)
Area GDP	5,620.639	4,360.677	2,987.963	4,972.767
	(5,493.355)	(4,372.221)	(3,021.635)	(4,780.787)
Area population	562.518	511.312	470.745	553.778
	(269.504)	(266.436)	(254.547)	(270.646)
Cameras per million residents	4.951	6.097	0.141	10.575
	(10.247)	(11.624)	(0.332)	(13.796)
Observations	796	308	103	137

Note: Observations at the procurement contract level. Standard deviations are reported below the mean. Administrative level of the contract is recorded as central government, provincial level, prefecture level and county level; the mean of an indicator of provincial or above level (provincial and central government) is shown. Local GDP is measured in millions of RMB, population in ten-thousand persons.

5 The impact of access to government data on AI firms

5.1 Empirical model and identification strategy

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 of the contract (i.e., surveillance capacity of the local public security agencies that issue the contracts).

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

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

The outcome variable, y_{it}, is the cumulative number of software releases by firm i up to the semi-year period t. The explanatory variables of interest are the interaction terms between a set of dummy variables, T_{it}, indicating semi-year time periods before or since firm i received its first public security AI contract, and $Data_i$, a dummy variable indicating whether the firm's first contract was data rich, as defined above.[28]

The coefficients on the interaction terms (i.e., on $\sum T_{it} \times Data_i$) 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. To account for time-varying sources of variation in software production common to all facial recognition firms (for example, government industrial policy promoting AI), we include time period fixed effects, α_t in all specifications. We also include firm fixed effects, γ_i, in all specifications, allowing us to control for all (observable or unobservable) time-invariant firm characteristics. Finally, in addition to estimating a parsimonious model without controls, we also estimate a model including a vector of pre-contract firm characteristics (X_i) interacted with time period fixed effects.[29] 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 ownership by a single mother firm.[30]

Our empirical strategy allows us to address important threats to identification. A particular concern is non-random assignment of contracts to firms. However, our triple differences identification strategy is not threatened by differential selection of firms into government contracts; nor is it threatened by selection into contracts with public security agencies. Rather, by exploiting variation within the set of firms receiving public security contracts, identification is threatened only by non-random selection of firms specifically into data-rich or data-scarce public security contracts. We account for fixed firm character-

[28] 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.

[29] Controls are firms' year of establishment, capitalization, and pre-contract software production.

[30] We cluster standard errors at the mother firm-level to be conservative; clustering standard errors at the firm level allows us to make even more precise inferences.

istics that may determine selection into data-rich contracts as well as software production by including a full set of firm fixed effects. We can test whether firms produced different amounts of software *prior* to receipt of a data-rich contract by testing whether β_{1T} differ from zero *prior* to contract receipt (that is, conducting a test of parallel pre-treatment trends). To address the possibility that *ex ante* firm characteristics shape selection into contracts and software production in a time-varying way, we control for firm characteristics interacted with time periods.

A second important concern is that contract characteristics other than data may affect software production. Many of these (such as a signal of a firm's connection to the government) are accounted for by differencing out the effects of data-scarce contracts, and we will also directly control for a contract's monetary size and a prefecture's GDP per capita interacted with time period fixed effects. In addition to including these controls, we will also present more direct evidence on the importance of data, as well as evidence against alternative mechanisms.

5.2 Results

5.2.1 Baseline estimates

We estimate our baseline specification, 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 2, we plot the coefficients β_{1T}, describing the *differential* cumulative 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 3, columns 1 to 4; columns 5 and 6 implement event study weighting adjustments, following Borusyak et al. (2017)). We show 95% confidence intervals for all coefficients, from models with and without controls ($\sum_T T_{it} X_i$). In Panel (a), one can see that receipt of a data-rich public security contract is associated with differentially more government software production than receipt of a data-scarce public security contract; in Panel (b), one observes a similar pattern regarding the *commercial* software production as well. In terms of magnitudes, we see that the receipt of a data-rich public security contract increases government software production by 2.9 and increases commercial software by 1.9 products over 3 years — on top of the effect of a data-scarce public security contract. Suggesting a causal interpretation, we find no evidence of pre-contract differences in software production levels or trends. The inclusion of controls for time-varying effects of firm characteristics has little effect on our findings.

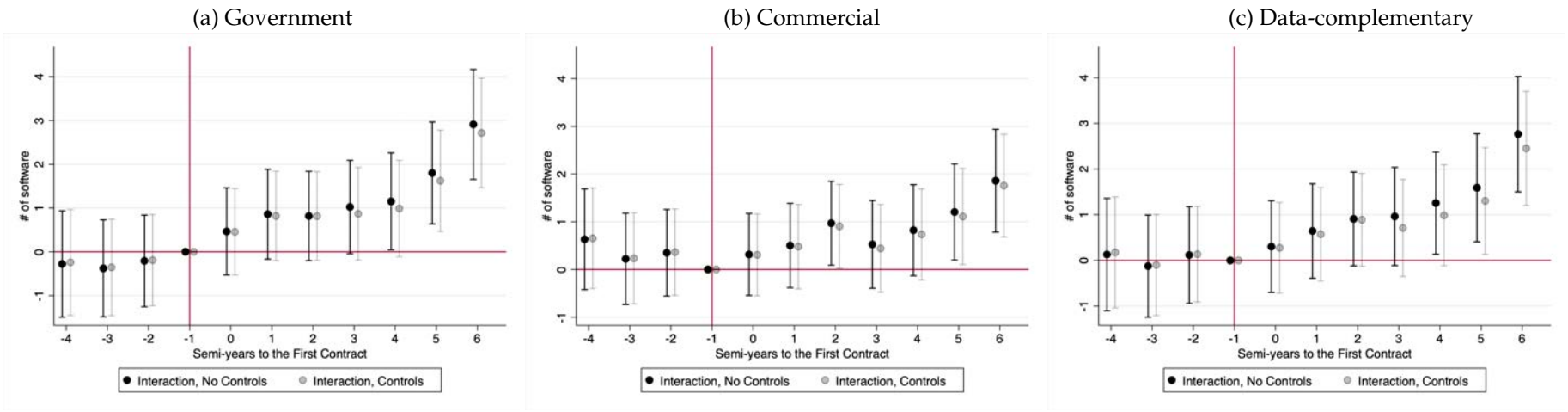


Figure 2: Differential cumulative software releases intended for government (left), for commercial (center), and for data-complementary uses (right) resulting from data-rich contracts, relative to data-scarce contracts, controlling for firm and time period fixed effects. Data rich contracts are defined as public security contracts in prefectures with above-median surveillance capacity. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.

Table 3: Data-rich public security contracts and AI software development

	Government	Commercial	Government	Commercial	Government	Commercial
	(1)	(2)	(3)	(4)	(5)	(6)
4 semiyears before	-0.177	-0.239	-0.182	-0.243	-0.322	-0.279
	(0.268)	(0.231)	(0.267)	(0.231)	(0.588)	(0.491)
3 semiyears before	-0.040	-0.180	-0.044	-0.183	-0.196	-0.231
	(0.264)	(0.228)	(0.262)	(0.227)	(0.437)	(0.365)
2 semiyears before	-0.002	-0.202	-0.004	-0.203	-0.087	-0.209
	(0.261)	(0.225)	(0.260)	(0.224)	(0.313)	(0.262)
Receiving 1st contract	0.750***	0.868***	0.680**	0.833***	0.880***	0.912***
	(0.279)	(0.239)	(0.277)	(0.239)	(0.327)	(0.273)
1 semiyear after	1.443***	1.663***	1.378***	1.630***	1.604***	1.663***
	(0.289)	(0.250)	(0.288)	(0.250)	(0.452)	(0.378)
2 semiyears after	2.243***	2.219***	2.106***	2.174***	2.424***	2.215***
	(0.301)	(0.258)	(0.300)	(0.258)	(0.603)	(0.503)
3 semiyears after	2.986***	3.122***	2.917***	3.087***	3.282***	3.119***
	(0.334)	(0.287)	(0.332)	(0.287)	(0.772)	(0.644)
4 semiyears after	3.984***	4.017***	3.910***	3.980***	4.330***	4.008***
	(0.360)	(0.309)	(0.358)	(0.308)	(0.942)	(0.786)
5 semiyears after	4.849***	4.857***	4.771***	4.817***	5.279***	4.883***
	(0.389)	(0.337)	(0.387)	(0.336)	(1.115)	(0.931)
6 semiyears after	5.595***	5.811***	5.511***	5.769***	6.036***	5.815***
	(0.444)	(0.378)	(0.441)	(0.378)	(1.297)	(1.081)
4 semiyears before × data-rich	-0.279	0.633	-0.243	0.653	-0.417	0.582
	(0.620)	(0.539)	(0.617)	(0.538)	(0.612)	(0.517)
3 semiyears before × data-rich	-0.379	0.222	-0.356	0.235	-0.424	0.198
	(0.565)	(0.488)	(0.562)	(0.487)	(0.557)	(0.468)
2 semiyears before × data-rich	-0.209	0.351	-0.192	0.362	-0.233	0.318
	(0.535)	(0.463)	(0.532)	(0.462)	(0.527)	(0.444)
Receiving 1st contract × data-rich	0.465	0.314	0.457	0.307	0.431	0.274
	(0.508)	(0.438)	(0.505)	(0.437)	(0.500)	(0.420)
1 semiyear after × data-rich	0.858	0.502	0.817	0.478	0.831	0.480
	(0.524)	(0.451)	(0.521)	(0.450)	(0.516)	(0.432)
2 semiyears after × data-rich	0.817	0.969**	0.814	0.904**	0.751	0.941**
	(0.520)	(0.449)	(0.518)	(0.449)	(0.514)	(0.432)
3 semiyears after × data-rich	1.023*	0.526	0.868	0.442	0.866	0.517
	(0.544)	(0.470)	(0.541)	(0.469)	(0.537)	(0.451)
4 semiyears after × data-rich	1.151**	0.823*	0.987*	0.735	1.007*	0.808*
	(0.565)	(0.487)	(0.562)	(0.486)	(0.558)	(0.468)
5 semiyears before × data-rich	1.800***	1.205**	1.623***	1.110**	1.628***	1.193**
	(0.594)	(0.515)	(0.591)	(0.514)	(0.587)	(0.495)
6 semiyears after × data-rich	2.911***	1.861***	2.715***	1.759***	2.761***	1.865***
	(0.642)	(0.550)	(0.638)	(0.549)	(0.634)	(0.529)
Controls	No	No	Yes	Yes	No	No
Event-study weighting	No	No	No	No	Yes	Yes

Notes: All regressions estimated on the sample of firms with first contracts with a public security agency. Baseline specification (Columns 1–2) controls for time period fixed effects and firm fixed effects. Columns 3–4 include controls for firms' pre-contract characteristics interacted with all semi-year indicators. Standard errors clustered at mother firm level are reported in parentheses. Columns 5–6 overweight (by 1000x) control groups (no contract firms) to address potential negative weighting issues in event studies (Borusyak et al., 2017). * significant at 10% ** significant at 5% *** significant at 1%.

Interpretation As discussed in Section 3, the results presented above indicate economies of scope in AI innovation arising from government data (or an algorithm trained with such data) being shared across commercial and government uses. In particular, the results imply that the benefits coming from access to government data outweigh any crowding-out of other resources from commercial software production within the firm, and that other inputs available in the private market must not be close substitutes for the government data firms are able to access. Importantly, our results are not merely capturing differentially less crowding 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.[31]

The results presented thus far do not appear to be the result of differential selection by firms into data-rich contracts. First, 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. Second, by differencing out the effects of data-scarce contracts, we account for time-varying selection into receiving a (generic) public security contract. Third, by controlling for the time-varying effects of firms' age and pre-contract software production, we address concerns about firms selecting into data-rich public security contracts as a function of their potential production growth. Finally, by controlling for the time-varying effects of firms' pre-contract capitalization, we account for selection into data-rich contracts on firms' potential benefit from the capital provided by a government contract. We find evidence of economies of scope arising from government data even including this full range of controls. In Sections 5.2.3 and 5.2.4, we provide further evidence of the importance of government data and address alternative interpretations.

5.2.2 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.[32] First, we assess the robustness of our results to restricting 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

[31] Appendix Figure A.10 plots the coefficients β_{2T} and $\beta_{1T} + \beta_{2T}$ for software production when a data-scarce and a data-rich public security contract were received, respectively.

[32] To present our findings in a concise manner, we report only a selection of coefficients that indicate software production two years before contract receipt among firms receiving data rich and data scarce contracts, respectively, as well as software production three years after contract receipt among firms receiving data rich and data scarce contracts, respectively. These coefficients allow one to observe any differential pre-trends, as well as the differential effects of data-rich contracts. The full set of coefficients are available from the authors, and show patterns fully in line with these selected coefficient estimates.

Table 4: Robustness

	Government	Commercial
	(1)	(2)
Panel A.1: Only major software releases		
4 semiyears before	-0.163	-0.197
	(0.265)	(0.227)
6 semiyears after	5.343***	5.719***
	(0.438)	(0.372)
4 semiyears before × data-rich	-0.292	0.607
	(0.612)	(0.529)
6 semiyears after × data-rich	3.107***	1.791***
	(0.633)	(0.540)
Panel B.1: LSTM categorization model configuration (timestep 10)		
4 semiyears before	-0.113	-0.310
	(0.275)	(0.324)
6 semiyears after	4.637***	4.948***
	(0.452)	(0.532)
4 semiyears before × data-rich	-0.328	0.521
	(0.638)	(0.760)
6 semiyears after × data-rich	2.516***	3.349***
	(0.658)	(0.775)
Panel C.1: LSTM categorization model threshold (60%)		
4 semiyears before	-0.139	-0.272
	(0.234)	(0.309)
6 semiyears after	3.465***	6.452***
	(0.389)	(0.508)
4 semiyears before × data-rich	-0.237	0.525
	(0.543)	(0.721)
6 semiyears after × data-rich	2.811***	2.349***
	(0.562)	(0.740)
Panel D.1: Time frame (full balanced panel)		
4 semiyears before	0.184	0.035
	(0.576)	(0.477)
6 semiyears after	5.634***	6.165***
	(0.728)	(0.597)
4 semiyears before × data-rich	-3.218	0.743
	(2.661)	(2.093)
6 semiyears after × data-rich	3.404***	2.048**
	(1.237)	(1.024)
Panel D.2: Time frame (extended time frame)		
5 semiyears before	-0.124	-0.204
	(0.274)	(0.236)
8 semiyears after	8.469***	6.986***
	(0.572)	(0.488)
5 semiyears before × data-rich	-0.342	0.269
	(0.686)	(0.597)
8 semiyears after × data-rich	3.793***	4.150***
	(0.756)	(0.648)

Panel E.1: Drop ambiguous public security agencies

4 semiyears before	-0.184	-0.260
	(0.270)	(0.230)
6 semiyears after	5.335***	5.916***
	(0.448)	(0.377)
4 semiyears before × data-rich	-0.375	0.625
	(0.649)	(0.557)
6 semiyears after × data-rich	3.222***	1.371**
	(0.659)	(0.558)

Notes: Specifications include full set of time indicators and interactions with data-rich contracts; only selected coefficient estimates are presented. Standard errors clustered at mother firm level are reported in parentheses. Panel A restricts the outcome software to only major releases (version X.0). Panel B varies the LSTM specification. Table 3, Columns 1-6 use the default LSTM specification with a timestep (phrase length) of 20, embedding size (number of dimensions in a vector to represent a phrase) of 32, and 32 nodes in the model. Panel B.1 presents results for the same model trained with a timestep of 10 instead. The full set of combinations of results with varied model parameters do not look qualitatively different. Table 3, Columns 1-6 use the default LSTM specification with a confidence threshold for the classification of software set at 50% (e.g. the model must be at least 50% confident that a given software is government software to be classified as "government"). Panel C.1 replicates the exercise setting the threshold to be higher, at 60%. Panel D.1 restricts the sample to firms that have non-missing observations during the entire time frame of 4 semi-years before and 6 semi-years after the initial contracts; Panel D.2 extends the time frame to 5 semi-years before and 8 semi-years after the initial contracts. Panel E.1 drops companies whose first contract is an ambiguous contract, or one that contains the keywords 'local government' ('人民政府') or 'government offices' ('政府办公室') which may be used for either public security or non-public security depending on interpretation. * significant at 10% ** significant at 5% *** significant at 1%.

(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.1 of Table 4).[33]

Second, we assess the robustness of our results to 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 B.1 of Table 4; see Panel A.1 and A.2 of Appendix Table A.4 for additional results).

Third, we evaluate the robustness of our results to adjustments of the LSTM classification threshold — the baseline specification sets the threshold at 50%. We re-categorize software using higher classification thresholds of 60% and 70% (requiring the algorithm to be even more confident that a software product belongs to a particular category for assignment), and these adjustments have no impact on our results (see Panel C.1 of Table 4; see Panel B.1 of Appendix Table A.4 for additional results).

Fourth, we can vary the time-frame studied: we examine wider windows of time around the receipt of the first contract; and, we consider a balanced panel of firms within a

[33] 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 Appendix Figure A.11 for the results in graphic form and Appendix Table A.5 for the results in regression form).

narrow window (studying a balanced panel over too long a window substantially reduces the sample size). These changes have no impact on our findings (see Panel D of Table 4).

Finally, we can vary the construction of the explanatory variable of interest, adjusting 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 bureaus could be meant to provide security services just for the government office building). This, too, has no impact on our results (see Panel E of Table 4).

5.2.3 Additional evidence of the importance of data as an input

Our proposed mechanism of economies of scope arising from government data suggests additional testable implications.

Data-complementary software Firms receiving access to unprecedented quantities of data may need to develop tools to manage that data (e.g., software supporting data storage). We next test whether firms receiving data-rich contracts differentially produce data-complementary software. Importantly, these data-complementary software products are *distinct* from the *AI* software studied above. In Figure 2, Panel (c), we present estimates from the same baseline specification, but now considering the outcome of data-complementary software products. One can see that data-complementary software production *differentially* increases after the receipt of a data-rich public security contract.[34] We find no evidence of pre-contract differences in data-complementary software production levels or trends, suggesting a causal effect of data-rich public security contracts.

Bidding patterns of procurement contracts Data-rich government contracts are more valuable to firms than data-scarce contracts. It is thus natural to test whether: *(i)* firms submit lower bids for data-rich contracts; and, *(ii)* more firms submit bids for data-rich contracts. While we do not have bidding information for all contracts, we use those contracts for which this information is available to estimate the relationship between bid values and local surveillance camera capacity at the time the contract was awarded, as well as the relationship between the number of bidders and local surveillance capacity. The patterns match what we expect (see Appendix Figure A.12): data-rich contracts are associated with lower bids — even controlling for bidding firm fixed effects (p-value $=$ 0.13) — and with more bidding firms (p-value $=$ 0.05).

[34] We find that data-complementary software increases after receipt of *both* data-scarce and data-rich contracts, with effects being significantly greater in the latter. All regression coefficients are presented Appendix Table A.5.

Alternative empirical definitions of data-richness of government contracts Finally, we consider two alternative empirical definitions of data-richness of government contracts. First, 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.[35] We thus consider an alternative definition of 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 (presented visually in Appendix Figure A.13, and in regression form in Table A.6).

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).[36] 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.

Second, we examine firms that produced video facial recognition AI software for the government following receipt of a public security contract: this software is the most data-intensive facial recognition AI software, presumably requiring access to the greatest quantity of government data.[37] We examine whether these firms also differentially produce more government and commercial software after receiving a data-rich public security contract. One can see in Appendix Figure A.15 that indeed they do. Moreover, we note that the magnitudes of the coefficients when considering the post-contract production of government video AI as an indicator of the data-richness of the contract are nearly double those using our other definitions, consistent with the idea that video AI software is particularly data-intensive.

A range of tests all point in the same direction: beyond other mechanisms through which government contracts may affect facial recognition software output, access to gov-

[35]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.

[36]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).

[37]Firms that produce video facial recognition AI for the government after receiving a data-rich public security contract also differentially produce more data-complementary software post-contract. See Appendix Figures A.14 (Panel B) and A.15.

ernment data plays a crucial role.

5.2.4 Evaluating alternative hypotheses

While a range of analyses suggest an important role for economies of scope arising from access to government data in shaping firms' production of AI software, it is important to consider alternative mechanisms. For a parsimonious presentation of the varied empirical exercises to come, in Figure 3, we plot regression coefficients and confidence intervals only for differential effects of data-rich contracts 3 years following contract receipt; the figure plots these estimates specification-by-specification. We also present more complete sets of estimates in Appendix Tables A.5 to A.7.

Differences in the terms and tasks under data-rich contracts One naturally wonders whether firms receiving data-rich public security contracts are engaged in similar work to firms receiving data-scarce public security contracts. We first examine whether differences in contractual terms may play a role in generating our results. To quantify the content of each public security contract, we calculate the vector distance between the language of the contract and a random sample of 500 non-public security contracts, using Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018). We control for this contract-specific distance measure interacted with time period fixed effects, and find that it cannot fully explain our results (see Figure 3 and Appendix Table A.7, Panel A).

We next compare the registered descriptions of firms' government software produced immediately following receipt of a data-rich or data-scarce public security contract. To quantify the content of each government software product description, we calculate the vector distance between the language of the government software descriptions and a random sample of 500 commercial software product descriptions, again using BERT. We test whether receipt of a data-rich contract differentially affects the government software produced by a firm (relative to receipt of a data-scarce contract); we find a very tight null result (government software descriptions change by around 1% of a standard deviation, with a p-value of 0.89). These results suggest that our findings are not driven by differences in the content of government software produced under data-rich and data-scarce contracts.

Learning by doing It is possible that data-rich contracts generate more AI software not because of the data they provide, but because of firms' opportunities for learning by doing under these contracts.[38] In particular, we evaluate two such possibilities. First,

[38]The crowding-in of government software production to commercial productivity that we document

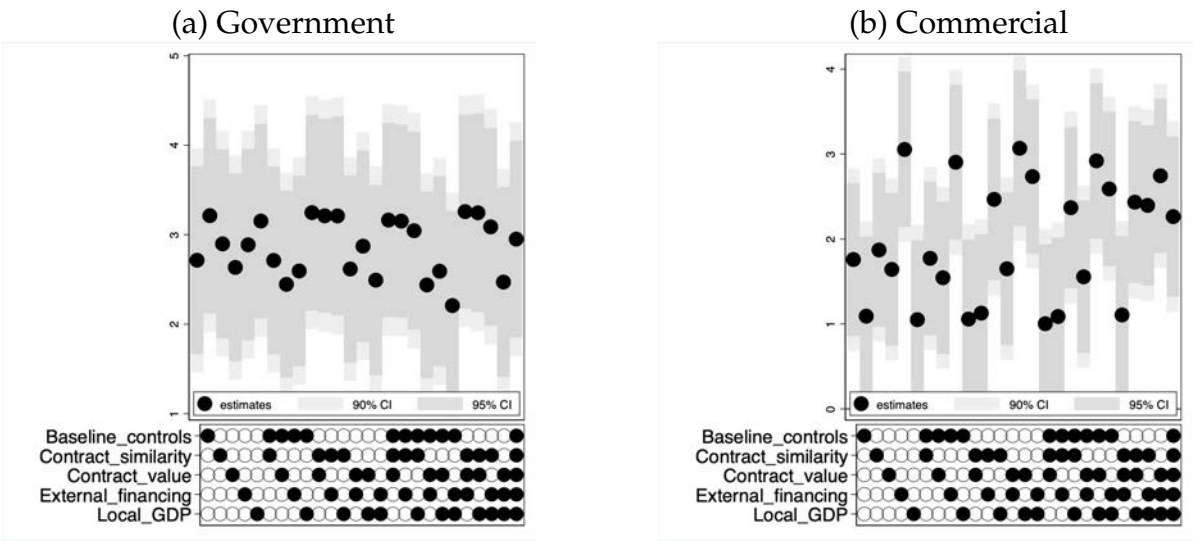


Figure 3: Software development intended for government (a) and commercial (b) use relative to the time of receiving initial procurement contract. Figure shows coefficient on the interaction for high surveillance capacity 3 years after public security contract receipt, controlling for firm and time period fixed effects and adding various controls. Solid dots indicate significance at the 10% level or better.

producing more government software might directly increase productivity in commercial software production. Second, due to differences in production processes, some types of government software may increase firm productivity in commercial software production more than others.

Note, however, that while learning by doing may be important in explaining the overall effects of contracts on software production, for it to explain our *differential* effects between data-rich and data-scarce contracts, it would have to be that the potential for learning (either due to the quantity or type of government software) was positively correlated with data-richness.[39]

Several pieces of evidence suggest that such systematically different learning by doing is not driving our main results. First, the potential for learning due to the quantity of software produced should presumably be stronger for firms with lower levels of production prior to the receipt of a contract. The time-varying control for pre-contract software production in the specification with controls (estimated above) allows us to (imperfectly) account for this. In addition, we estimate our baseline specification, but now including time-varying controls for pre-contract government software production, software production in the corresponding category, or software production in the opposite category

is both immediate and persistent, which differs from the learning by doing patterns observed in other contexts. For example, a long lag before the crowd-in takes place (e.g., engineers who worked at NASA transferred knowledge to the civilian aviation industry only years later), or with short duration (e.g., managers in the garment sector who learned to improve their managerial skills, but experienced quick decay in learning (Adhvaryu et al., 2019)).

[39]Some forms of what could be thought of as learning by doing are precisely part of the mechanism that we are trying to capture. For example, we expect improved algorithmic performance as a result of more predictions made on larger datasets. If algorithms, as opposed to data, are sharable across uses we would also label this as economies of scope arising from government data.

(e.g., controlling for government software production when examining commercial software production as outcomes). These controls only slightly reduce the effect of a data-rich contract (see Appendix Table A.7, Panel B). Finally, systematic software differences should appear in software descriptions. However, we have shown above that the description of government software produced following the receipt of a data-rich public security contract is very similar to the software produced after the receipt of a data-scarce one.

Government contracts as sources of capital Another important consideration is that contracts may affect firms' software production through the provision of capital. We attempted to account for this channel above by differencing out the impact of "data-scarce" contracts and by controlling for the time-varying effects of firms' pre-contract capitalization, but we can also address this concern in two other ways. First, we can directly control for the monetary value of the contract interacted with time period fixed effects (formally $\sum_T T_{it} value_i$). We add these interactions to our baseline specification and find that they do not affect our results (see Figure 3 and Appendix Table A.7, Panel A). Second, we add to our baseline specification interactions between a firm's pre-contract amount of external financing and the full set of time period fixed effects (formally $\sum_T T_{it} \times financing_i$). Again, they have no impact on our results (see Figure 3 and Appendix Table A.7, Panel A).

Government contracts as signals It is also possible that receipt of a data-rich contract may function as a signal of firm quality or potential: 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. To test whether the differential signaling 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 signaling value of these first contracts should be lower (mother firm quality is already observed), while access to data remains potentially extremely valuable. In Appendix Table A.7, Panel C, one can see that within this sample of firms belonging to a mother firm that has *already* received a government contract through a different subsidiary, there is still a significant differential effect of receiving a data-rich contract on both government and commercial software production.

Different commercial opportunities associated with data-rich contracts A last important set of concerns is that contracts with governments in prefectures with high surveillance capacity may offer different commercial opportunities for reasons other than the

additional data to which firms gain access. First, high-surveillance prefectures may also be richer commercial markets; a contract with a local government in a richer prefecture could affect software production. To evaluate this possibility, we control for the GDP per capita of the administrative unit where a firm's first government contract was issued, interacted with time period fixed effects (formally $\sum_T T_{it} \times market_i$). Adding these interactions to our baseline specification does not affect our results (see Figure 3 and Appendix Table A.7, Panel A). A second possibility is that contracts with two very specific high-surveillance prefectures may disproportionately affect our results: Beijing and Shanghai. Contracts with these powerful local governments may offer a range of political and economic opportunities that go beyond access to data. To rule out the possibility that our findings are distorted 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 (see Appendix Table A.7, Panel D). A third possibility is that contracts with a firm's home-province government may give the firm some commercial advantage, beyond the effects of data. 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 (see Appendix Table A.7, Panel D).

Our empirical results thus paint a clear picture: after receiving government contracts that provide them with greater access to government data, firms are able to use that data (or transferable algorithm trained with it) to develop not only government software products, but also commercial software products. This is possible due to the economies of scope arising from government data, rather than other mechanisms.

6 A directed technical change model with data as an input

In our empirical analysis of Section 5, we have shown the *firm-level* consequences of access to government data: an increase in government data available to AI firms increases their government and commercial innovation. We next ask: what are the trade-offs presented by states' AI procurement and data provision policies in the age of data-intensive innovation? In order to answer this question, we now build a directed technical change model (Acemoglu, 2002) with data as an input and economies of scope generated by government data. In Section 7, we use this model to analyze the positive and normative implications of such state policies.

Model overview We model an economy in which firms innovate to develop and sup-

ply differentiated varieties of government and commercial (private) software — which require data in production — as well as other, non-software, varieties — which do not. Commercial software and non-software varieties are intermediate inputs into the production of a final good. A representative household consumes the final good and owns all firms. Government software varieties are purchased by the state as intermediate inputs to produce a government good. To be concrete and link it to our empirical setting, we refer to this government good as "surveillance."

As in Section 3, we assume that government data can be shared across uses within the firm. Specifically, government data is necessary for producing government software and the same data can simultaneously be used for producing commercial software — where it is not necessary and is instead a gross substitute with private data. Government data is supplied by the state and is produced as a by-product of surveillance. Private data is supplied by a representative firm as a by-product of all private transactions in the economy as measured by total output of the final good.[40] Furthermore, while both types of data are excludable, we assume that only private data can be purchased in the market. In contrast, as in Section 3, government data can only be accessed by obtaining a contract for producing government software varieties for the state.

The state chooses a policy that involves: a level of expenditures on surveillance (which determines the amount of government data produced), an amount of government data supplied to firms that obtain a contract to produce government software varieties, and the levels of lump sum taxes of, and transfers to, households. Given a state policy, potential entrants can choose to innovate on and supply new varieties of government software, commercial software, both types of software, or only non-software varieties. Firms will innovate and enter such that, in a balanced growth path equilibrium, all sectors grow at the same rate, and returns to innovation are equalized across sectors. We next describe this economy formally.

Goods production Consider an economy with three intermediate good sectors producing: commercial (private) software Y_c, government software Y_g, and other non-software products Y_z. Within each sector i, there is a measure N_i of differentiated product varieties j of quality $q_i(j)$. A representative sectoral firm has production technology:

$$Y_i = \frac{1}{1-\frac{1}{\chi}} \int_0^{N_i} q_i(j)^{1-\frac{1}{\chi}} dj. \tag{1}$$

[40] This corresponds, for instance, to information collected from consumers when performing online transactions. More broadly, this reflects the pattern that private data is produced as a result of commercial innovation that generates more private consumption. With this setup though, we are ignoring interesting issues regarding how to allocate private data property rights between firms and consumers.

We assume the firm is competitive and maximizes static profits taking sectoral prices p_i and product variety prices $p_i(j)$ as given. This gives inverse demand schedules:

$$p_i(j) = p_i q_i(j)^{-\frac{1}{\chi}}. \tag{2}$$

A representative firm then combines private software and non-software to produce a final good Y using a CES aggregator:

$$Y = \left[a Y_z^{\frac{\epsilon-1}{\epsilon}} + (1-a) Y_c^{\frac{\epsilon-1}{\epsilon}} \right]^{\frac{\epsilon}{\epsilon-1}}. \tag{3}$$

We again assume the firm is competitive and maximizes profits given prices p_c and p_z, and the price of Y which we normalize to 1. This implies that prices satisfy:

$$1 = \left((a)^{\epsilon} (p_z)^{1-\epsilon} + (1-a)^{\epsilon} (p_c)^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}}. \tag{4}$$

Innovators A software variety j is supplied by a monopolist "innovator." As in Section 3, we assume that producing software of a higher quality is data-intensive.[41] Dropping the j index for notational convenience, government software production uses government data d_g and intermediate goods x_g to produce a variety of quality q_g. Commercial software production uses both government and private data, d_g and d_p, as well as intermediates x_c to produce a variety of quality q_c.[42]

Specifically, we assume that the firms may produce government and commercial software using the following technologies (a special case of those in Section 3):

$$q_g(d_g, x_g) = (d_g)^{\beta} x_g^{1-\beta} \tag{5}$$

$$q_c(d_g, d_p, x_c) = \left(\alpha d_g^{\frac{\gamma-1}{\gamma}} + (1-\alpha) d_p^{\frac{\gamma-1}{\gamma}} \right)^{\frac{\gamma}{\gamma-1}\beta} x_c^{1-\beta}, \tag{6}$$

where $\alpha < 1$ governs the relative productivity of government vis-à-vis private data, and $\gamma > 1$ describes their gross substitutability in commercial software production.[43] With this specification, α is a key parameter governing the strength of economies of scope generated by government data.

Next, we consider the profit maximization problem for a software variety of quality q. We assume that private data can be purchased in the market at price p_d. Moreover, we assume that intermediate goods x_g, x_c cost ϕ units of the final good (whose price is

[41] For example, one measure of quality of AI facial recognition software is prediction accuracy. This is higher when larger datasets are used in training the AI algorithms.

[42] For simplicity, we assume no private data is used in government software production. This is not necessary though and does not affect our qualitative results. But it would matter for our numerical examples.

[43] The assumption of gross substitutability is important because, as will be seen below, it allows innovators to produce commercial software even without access to government data.

normalized to 1) and that varieties never depreciate.[44]

These assumptions, together with demand schedules for a variety having constant elasticity χ imply that, for any sectoral price p_i and government data d_g, the flow of profits from a variety are:

$$\Pi_g(d_g, p_g) = \max_{x_g} p_g q_g(d_g, x_g)^{1-\frac{1}{\chi}} - \phi x_g, \tag{7}$$

$$\Pi_c(d_g, p_c, p_d) = \max_{x_c, d_p} p_c q_c(d_g, d_p, x_c)^{1-\frac{1}{\chi}} - \phi x_c - p_d d_p, \tag{8}$$

and the corresponding input demand schedules are $d_p(d_g, p_c, p_d), x_c(d_g, p_c, p_d), x_g(d_g, p_g)$.

Next, we describe how new varieties are introduced. We assume that innovators can invest 1 unit of the final consumption good in R&D in order to produce μ_i new varieties in sector i — thus becoming the monopolist supplier of those varieties forever.[45] Then, given total R&D spending R_i for sector i, new varieties accumulate according to:

$$\dot{N}_i = \mu_i R_i. \tag{9}$$

The entry decision is somewhat nuanced due to the fact that government data can be shared across uses and that there is no market for such data. We assume the following sequence of events takes place. A software innovator can first decide whether to attempt to obtain a government contract or not by paying a cost F. If the innovator decides not to make an attempt, it can choose to introduce a new commercial software variety without access to government data ($d_g = 0$). If it decides to make an attempt, it obtains a government contract with probability λ. The contract commits the innovator to produce a new government software variety and provides the innovator with access to a fixed quantity of government data $\bar{d}_g$. The innovator can then choose to also introduce a new commercial software variety using government data in its production. Finally, if the innovator does not obtain the government contract, it can again choose to introduce a new commercial software variety without access to government data.

We consider a balanced growth path (BGP) with constant interest rate r and free-entry of innovators. This implies that the expected present discounted value of profits net of the unit cost of R&D investment must be zero for both government and commercial software innovators. Given these assumptions and setting $\mu_g = \mu_c = 1$, a BGP equilibrium with

[44] As in Acemoglu (2002), if varieties depreciate slowly, this would not change the balanced-growth path equilibrium — which will be our focus — but only the transitional dynamics.

[45] We use a "lab equipment model" of innovation which emphasizes reproducible resources, like electricity and hardware, that play an important role in the context of AI. One could also incorporate researchers too, as in a "knowledge-based R&D model" (Acemoglu, 1998), without changing the qualitative implications.

both types of software firms present requires:[46]

$$F = \lambda \left(\frac{\Pi_g(\bar{d}_g, p_g)}{r} - 1 + \max \left\{ \frac{\Pi_c(\bar{d}_g, p_c, p_d)}{r} - 1, 0 \right\} \right) + (1 - \lambda) \max \left\{ \frac{\Pi_c(0, p_c, p_d)}{r} - 1, 0 \right\}, \tag{10}$$

$$1 = \frac{\Pi_c(0, p_c, p_d)}{r}. \tag{11}$$

Finally, for non-software innovators which do not require data as an input, the R&D investment yields new varieties with quality $q_z = x_z^{1-\beta}$, where x_z is again intermediate goods. This results in profits:

$$\Pi_z(p_z) = \max_{x_z} p_z q_z^{1-\frac{1}{\chi}} - \phi x_z. \tag{12}$$

The free-entry condition for non-software innovators is then:

$$1 = \mu_z \frac{\Pi_z(p_z)}{r}. \tag{13}$$

Representative household We assume the existence of a representative household with CRRA flow utility $u(C) = \frac{C^{1-\theta}}{1-\theta}$, where C is consumption of final goods and θ is the inverse of the intertemporal elasticity of substitution. Then, given discount rate ρ, the present discounted utility is:

$$\int_0^\infty e^{-\rho t} u(C_t) dt \tag{14}$$

In Section 7, we introduce extensions to the household utility that allow: *(i)* the government good to affect utility, either positively or negatively; and *(ii)* data collection itself to impose a cost.

The household maximizes utility subject to the budget constraint:

$$C_t + \dot{A}_t \leq A_t r_t + \Pi_t - T_t, \tag{15}$$

where A_t are assets, Π_t are profits coming from all firms, and T_t are taxes.

Data supply and the state The state purchases the government software aggregate Y_g at price p_g in order produce surveillance G with linear technology $G = Y_g$. It sets lump sum taxes T on households so that budget balance holds at each time:

$$p_g G = T. \tag{16}$$

Aggregate government data D_g is produced as a by-product of government surveil-

[46] As seen in equation (10), we abstract from the possibility that government production crowds-out resources from commercial production. We do so for simplicity and because, empirically, we have shown that the sharability of government data dominates, resulting in overall crowding-in.

lance: specifically, one unit of surveillance, G, produces κ_g units of government data.[47] Then, given a measure N_g of government software innovators and a dataset available to them $\bar{d}_g$, we have that:

$$N_g \bar{d}_g = D_g = \kappa_g G. \tag{17}$$

As can be seen in equation (17), we assume that government data is not sharable *across* firms. We do so for two reasons. First, to conceptually focus on the positive and normative implications of the sharability of government data across uses *within* a firm (the consequences of non-rival private data across firms have been studied by, e.g., Jones and Tonetti, 2018). Second, because in our empirical setting this seems to be the more relevant case. While sharing government data across firms may be feasible from a technological standpoint, we observe local governments collecting their own surveillance data and contracting with specific firms to analyze it, thus implicitly excluding other firms from its use. We note though, that allowing government data to be sharable across firms as well would magnify the overall importance of government data in our model.

We are now ready to formally define a state policy. Because we will consider a balanced growth path, we find it more useful to define the policy in terms of variables that are stationary. In particular, we divide the level of government software expenditures for surveillance and lump sum taxes by the level of private output.

Definition 1 (State policy) *A state policy is a dataset available to government software innovators $\bar{d}_g$, government software expenditures for surveillance purposes relative to final good output $p_g G/Y$, and lump sum taxes relative to final good output T/Y that satisfy equations (16) and (17).*

Finally, we complete the description of the economy's environment with the production of private data. A representative firm produces D_p by "mining" data out of private transactions as measured by total private output Y.[48] Suppose it can mine $\kappa_p Y$ units of data out of Y, then the supply of private data is:[49]

$$D_p = \kappa_p Y. \tag{18}$$

[47] In our empirical context, this government data would correspond to the faces on video feeds captured by street cameras. These are themselves produced as a by-product of surveillance and public security provision, the activites carried out by government units.

[48] Requiring other inputs in private data production function and having a less than perfectly elastic data supply does not matter much qualitatively; though changes in the supply elasticity would affect the results of our subsequent numerical exercises.

[49] Note that this firm will be making positive profits in equilibrium. One interpretation of these profits is that they are rents from ownership of a fixed factor that is needed in order to mine private data. For example, in reality, the fixed factor could be the "land" on which data centers are built.

Equilibrium We now consider a BGP equilibrium where all variables grow at constant rate η. We denote $\tilde{N}_c$ as the total number of commercial software varieties produced by firms without a government contract, N_g as the number of commercial software varieties produced by firms with a government contract (which is also the number of government software varieties), and N_z as the number of non-software varieties.[50]

Definition 2 (BGP equilibrium) *Given a state policy $\{\bar{d}_g, p_g G/Y, T/Y\}$, a balanced-growth path equilibrium is a set of prices $\{p_c, p_z, p_g, p_d, r\}$, relative varieties $\tilde{N}_c/N_z$ and N_g/N_z, and growth rate η such that firms and households are optimizing, there is free-entry of innovators, and all markets clear.*

Because we endogeneize the production of data and new software varieties, it is possible that, for some parameterizations, no BGP equilibrium exists with entry of both types of software firms: i.e., those producing commercial software alone and those producing both government and the commercial software. Proposition 1 in Appendix A.1 lays out sufficient conditions for a BGP equilibrium to exist and be unique where all types of firms are present.[51]

We now formally define two objects that will be of interest next. The first is the economy's BGP growth rate η, which equals the rate of innovation in any sector i:

$$\eta = \frac{\dot{N}_i}{N_i}. \tag{19}$$

The second is the bias of private innovation towards data-intensive software, which we define as commercial software varieties relative to non-software varieties along the BGP:

$$n_c = \frac{N_c}{N_z}, \tag{20}$$

where N_c is an output-weighted average of commercial software varieties $N_c \equiv \tilde{N}_c\omega + N_g(1-\omega)$, with $\omega = \frac{q_c(0,p_c,d_p)^{1-\frac{1}{\chi}}}{q_c(0,p_c,d_p)^{1-\frac{1}{\chi}} + q_c(\bar{d}_g,p_c,d_p)^{1-\frac{1}{\chi}}}$.

[50] We denote by $\tilde{N}_c$ the *subset* of commercial software varieties produced by firms using *only* private data; we reserve the notation N_c to capture all types of commercial software varieties (as discussed below).

[51] This is the empirically relevant equilibrium: most AI firms produce commercial software *without* access to government data.

7 State policies and trade-offs in the age of data-intensive innovation

States' procurement of data-intensive technologies, as demonstrated in China's facial recognition AI sector, involves two policy dimensions: first, how much of the data-intensive technology to demand, and second, how much data to collect and provide to firms that produce it. In this section, we analyze the positive and normative implications of these two state policy dimensions. Each policy choice presents important trade-offs beyond purely economic considerations. The state's use of data-intensive technologies may harm citizens by infringing on their civil liberties (e.g., in surveillance states and informational autocracies); we study this in Section 7.1. Even when states' use of data-intensive technologies benefits citizens, they may dislike states' data collection and provision to firms (e.g., due to concerns about privacy violations); we study this in Section 7.2.

7.1 States' demand for surveillance AI

All states engage in citizen monitoring and surveillance to ensure public security. In the modern world, state monitoring is likely to involve substantially greater data collection and data analysis — particularly using AI. Some of this data-intensive surveillance may be beneficial to citizens (e.g., due to the preservation of public order), while some can impose direct harm (e.g., due to severe infringement of civil liberties). At the extreme are autocratic states that aim to monitor and control their populations to maintain power (Guriev and Treisman, 2019). Indeed, AI has been described by the *Wall Street Journal* as part of the "autocrat's new tool kit."[52] China is a prototypical example of this phenomenon, leading the world in surveillance capacity: there will be around 560 million public surveillance cameras installed in China by 2021, versus approximately 85 million in the US.[53]

We now use our model to analyze the consequences of a state policy involving larger government software expenditures for surveillance purposes relative to final good output $p_g G/Y$.[54] We begin with a purely positive analysis and then discuss normative implications and trade-offs. We conclude by connecting these results back to our empirical context and provide a numerical illustration of the forces at play.

[52]Source: `https://on.wsj.com/2H1sIgu`.

[53]Source: `https://on.wsj.com/2UOuuIJ`.

[54]Note that analogous results will arise from other government expenditure on software, such as public health and mapping.

Positive implications The next theorem shows the conditions under which an increase in $p_g G/Y$ causes an increase the economy's growth rate and biases the direction of private innovation towards data-intensive software.

Theorem 1 *Assume the sufficient conditions in Proposition 1 for a unique BGP equilibrium to exist hold. Then, an increase in the BGP's government surveillance expenditures relative to final good output ($p_g G/Y$) will increase the rate of innovation (η). Moreover, if relative demand for software is sufficiently elastic ($\epsilon \geq \frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}$), it will also bias private innovation towards data-intensive software (increase n_c).*

Proof. See Appendix A.2. ■

Beyond the formal proof, we also provide an intuitive discussion of the theorem in Appendix A.2. In brief, the higher state demand for surveillance increases both the equilibrium price p_g and the supply of government data to firms $\bar{d}_g$ which is produced as a by-product. This drives up the profits earned by firms using government data for government and commercial software development (due to economies of scope). Under free entry, this then increases the return on investment (r) and, in turn, induces higher R&D spending and increases the rate of innovation on the BGP. Moreover, in equilibrium, innovators must be indifferent among developing software varieties using government data, developing commercial software without using government data, and developing non-software varieties. The necessary price adjustments for such indifference imply that commercial software sells at lower prices in the new equilibrium. If relative demand is sufficiently elastic ($\epsilon \geq \frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}$), this implies that the new entry of commercial software innovators will be sufficient to bias private innovation towards data-intensive software.

Theorem 1 and our empirical evidence on economies of scope together suggest a potential alignment between surveillance states and data-intensive innovation. Greater purchases of government software and surveillance production will not only increase the state's political control, but also produce government data (as a by-product) that fuels growth and commercial data-intensive innovation when there are economies of scope.

Normative implications The choice of surveillance and public security spending involves not only economic considerations but also political. Thus, to be able to go beyond analyzing purely economic trade-offs, we consider an extension of our baseline model where the flow of household utility is:

$$\left(G^{\psi}C - \frac{\delta}{1+\psi}G^{1+\psi}\right)^{1-\tilde{\theta}} \frac{1}{1-\tilde{\theta}}. \tag{21}$$

Beyond being consistent with a BGP, this formulation captures in a stylized way both the potential benefits of higher surveillance and public security provision for households — due to, for example, lower crime or terrorism — and their potential costs due to the infringement of civil liberties. Formally, note that when $\delta G > \psi C$ the marginal utility of G is negative.

With this in mind, we now analyze the BGP welfare implications of a state policy that increases $p_g G/Y$. Assuming that utility is bounded, the present discounted utility of the representative household is:[55]

$$U = \underbrace{\frac{1}{\rho - (1-\tilde{\theta})(1+\psi)\eta}}_{\text{Growth effect}} \left(\underbrace{\frac{C}{Y}}_{\text{Consumption effect}} - \underbrace{\frac{\delta}{1+\psi}\frac{G}{Y}}_{\text{Surveillance effect}} \right)^{1-\tilde{\theta}} \underbrace{\left(\frac{G}{Y}\right)^{\psi(1-\tilde{\theta})}}_{\text{Surveillance effect}} \frac{1}{1-\tilde{\theta}}.$$

A state policy that increases in $p_g G/Y$ thus has the following household welfare implications. The increase in η shown in Theorem 1 leads to a direct positive effect on welfare, since the growth rate of consumption of the private and government goods is higher. But, there are two potentially offsetting forces. First, the proof and discussion of Theorem 1 show that G/Y increases as well, which will have a negative effect on utility when δ is high enough so that $\delta G/Y > \psi C/Y$ (namely, a negative surveillance effect). Second, from the aggregate resource constraint (shown below), we see that the private consumption to output ratio C/Y may decrease due to crowd-out of resources used in creating new varieties (i.e., innovation) and as intermediate inputs (i.e., production):

$$\frac{C}{Y} = 1 - \underbrace{\left(\left(2 + \frac{F}{\lambda}\right) \frac{\dot{N}_g}{Y} + \frac{\dot{\tilde{N}}_c}{Y} + \frac{1}{\mu_z}\frac{\dot{N}_z}{Y} \right)}_{\text{Resources used in innovation}} - \underbrace{\frac{\chi - 1}{\chi}(1-\beta)\left(1 + \frac{p_g G}{Y}\right)}_{\text{Resources used in production}}.$$

Discussion and numerical example Going back to our empirical context, the above positive and normative results imply that the demand for AI to support a surveillance state like China may incidentally increase innovation and promote economic growth. Yet, it will simultaneously bias the direction of innovation towards government and commercial AI, therefore potentially reducing citizen welfare due to both the crowd-out of resources from consumption and the infringement of civil liberties when surveillance is excessive.

[55] We normalize the initial output level to 1. The only difference from our baseline model is that the BGP Euler equation becomes $-\tilde{\theta}\frac{G_t^{\psi} C_t \frac{\dot{C}_t}{C_t} + \left(\psi G_t^{\psi} C_t - \delta (G_t)^{1+\psi}\right)\frac{\dot{G}_t}{G_t}}{G_t^{\psi} C_t - \frac{\delta}{1+\psi}(G_t)^{1+\psi}} + \psi \frac{\dot{G}_t}{G_t} = \frac{\dot{\lambda}_t}{\lambda_t}$. Assuming all variables grow at the BGP rate η gives the Euler condition $r = \rho + (\tilde{\theta}(1+\psi) - \psi)\eta$. Thus, all of our previous results apply without change to this economy by re-interpreting θ as being equal to $\tilde{\theta}(1+\psi) - \psi$.

To illustrate these forces, we now report the results of a simple numerical example. Our objective is not to provide a comprehensive quantitative evaluation but to give a sense of the implications of our empirical evidence on economies of scope for the normative trade-offs at play. We fully acknowledge that this exercise is speculative: while we have a good quantitative sense about the strength of economies of scope (presented in Section 5), we have a large degree of uncertainty about other key parameters in the model, such as the elasticity of substitution across sectors (ε), the share of data versus other inputs in production (β), or the extent to which surveillance affects utility (ψ, δ).

To make the analysis transparent, we first externally fix certain parameters (like θ, β and ε) and then calibrate the remaining ones so that in a baseline specification: *(i)* the economy is symmetric in the sense that the direction of innovation is unbiased ($\frac{\tilde{N}_c}{\tilde{N}_z} = \frac{N_g}{N_z} = 1$), all sectors have an identical share ($\frac{p_c Y_c}{Y+p_g G} = \frac{p_z Y_z}{Y+p_g G} = \frac{p_g G}{Y+p_g G} = 1/3$), and private and government data demands are identical ($\tilde{d}_g = d_p(\tilde{d}_g, p_c, p_d)$); and, *(ii)* economies of scope (as governed by α) are consistent with our benchmark estimates from Section 5 (to be precise, the relative elasticity of commercial to government software production of around two-thirds implies $\alpha = 0.8$).[56] Then, we vary the level of government surveillance spending $p_g G/Y$ from this benchmark parameterization and compute the BGP rate and bias of innovation as well as household welfare.

The left panel of Figure 4 shows the positive implications of the policy for the case with $\delta = \psi = 0$, thus illustrating Theorem 1. For example, the bias of private innovation towards data-intensive software falls from 1 to 0.6 when moving from a government spending share of $1/3$ (the symmetric economy) to an economy with no government spending, whereas the growth rate falls from 0.06 to 0.02.

The right panel shows the normative implications for three extreme cases: *(i)* the benchmark case with only economic trade-offs at play ($\delta = \psi = 0$); *(ii)* a case where the direct effect of surveillance on household utility is always negative ($\delta = 0.1, \psi = 0$); and *(iii)* a case where it is always positive ($\delta = 0, \psi = 0.2$). In all cases welfare is expressed in consumption equivalent deviations from the symmetric economy with $\delta = \psi = 0$. Given our parameterization, we find that a benevolent state seeking to maximize household welfare would choose to have a strictly positive government spending when $\delta = 0$ (and *a fortiori* when ψ = 0.2) but not when $\delta = 0.1$. Therefore, depending on the nature of surveillance and how strong resource crowd-out is, a benevolent state may choose to spend little in data-intensive surveillance technologies.

In contrast, a non-benevolent state may choose a higher level of spending on surveillance than that which maximizes citizens' welfare (e.g., because surveillance increases

[56] See Appendix B.2 for a more detailed description of the calibration.

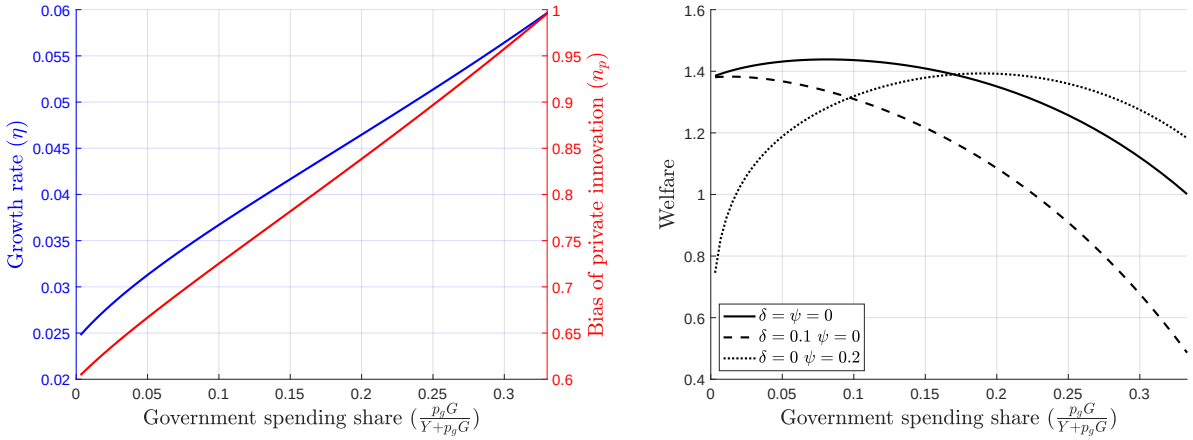


Figure 4: Positive (left panel) and normative (right panel) implications of changes in government spending share $\frac{p_g G}{Y+p_g G}$. Left panel is for benchmark parameterization under the symmetric economy (see Appendix B.2) and $\delta = \psi = 0$. Welfare in the right panel is measured in consumption equivalent deviations from the symmetric economy with $\delta = \psi = 0$.

control over their citizens and the likelihood an autocrat remains in power). In so doing, it will distort innovation, biasing it towards data-intensive software, and increase the growth rate — but at the expense of citizen's welfare. Even when the only cost is resource crowd-out from consumption (the case with $\delta = \psi = 0$), welfare declines by about 40 percent when moving from the household welfare maximizing level of spending of about 0.1 to the symmetric economy's spending share of $1/3$. Moreover, the welfare losses from excessive state surveillance may be much larger when it also imposes a non-pecuniary cost on citizens.

7.2 Government data provision as a form of innovation policy

In addition to choosing how much of the data-intensive technology to purchase, states must also decide how much government data to collect and share with firms for which government data is a key input into innovation. Indeed, even if the government good produced with this technology was not directly valued by the household, a state could choose to provide government data to firms as a form of innovation policy. This would be similar to other innovation and industrial policies that often entail a direct provision of key production inputs to private firms — including, for example, transportation or electric power infrastructure as well as public services that increase worker productivity like education or health.[57] However, one important aspect that distinguishes government

[57]For example, see Barro (1990) for a canonical endogenous growth model with government provided goods as an input in production.

data from these other inputs is that individuals often express particular discomfort when their data is collected and shared by the state.[58] We now turn to studying the positive and normative implications of government data provision to firms. We ask whether providing government data to innovating firms may be justified in the age of data-intensive innovation, taking into consideration citizens' privacy concerns.

Positive implications We begin by establishing an equivalence between a state policy choice that increases the amount of government data provided to firms ($\bar{d}_g$) and the policy choice from Theorem 1 which considered instead an increase in surveillance spending $p_g G/Y$.

Corollary 1 *Under the conditions of Theorem 1, a state policy that increases government data provision to firms ($\bar{d}_g$) is equivalent to a policy that increases government spending $p_g G/Y$ in terms of their implications for the BGP growth rate (η) and the bias of private innovation (n_p).*

The proof of the result is in Appendix A.2. From a positive perspective, the equivalence with Theorem 1 shows that providing more government data to innovators producing government software varieties increases the economy's growth rate and biases the direction of private innovation towards data-intensive software. Moreover, this equivalence implies that, even when pursuing a different objective, a state's demand for AI (as in Theorem 1) may also have *incidental* industrial policy elements, echoing the arguments of Rodrik (2007) that all policies, whether intended or not, can be considered as industrial policies. Finally, we note that a policy that provides government data to all firms — not just those contracting with the state as in our empirical context — would further foster innovation in our model.

Normative implications We showed above that increases in $\bar{d}_g$ can lead to a higher growth rate η. Yet, there is no reason for the state to introduce a policy that increases the growth rate *per se*. The appropriate objective for a benevolent state is to maximize household utility. Given this discussion, we now consider a second-best problem where the state chooses the level of government data provision to maximize household welfare.[59]

Similar to Section 7.1, we extend our baseline model to go beyond economic trade-offs alone. Specifically, we let the flow household utility be:

$$\left(C - \varphi D_g\right)^{1-\theta} \frac{1}{1-\theta},$$

[58] Source: "Customer Data: Designing for Transparency and Trust", `https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust`.

[59] It is a second-best problem because of distortions coming from the monopoly power of innovators in the decentralized equilibrium.

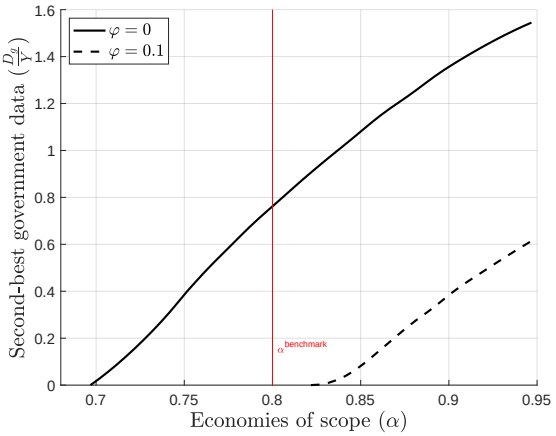


Figure 5: Second-best government data provision $(\frac{D_g}{Y})$ as a function of the strength of economies of scope (α) and privacy concerns (φ). Red line corresponds to our benchmark estimate for α (see Appendix B.2 for details).

where $\varphi > 0$ captures the idea that the collection and sharing of government data causes a direct disutility to the household. In a BGP, the present discounted utility of the representative household is then:

$$U = \underbrace{\frac{1}{\rho - (1-\theta)\eta}}_{\text{Growth effect}} \left(\underbrace{\frac{C}{Y}}_{\text{Consumption effect}} - \underbrace{\varphi \frac{D_g}{Y}}_{\text{Privacy effect}} \right)^{1-\theta} \frac{1}{1-\theta}.$$

The above expression implies that a benevolent planner choosing the optimal level of government data provision $\bar{d}_g$ will trade off the welfare gains from higher economic growth with the losses coming from the crowding-out of resources from consumption (as in Section 7.1) and citizens' disutility due to privacy concerns from government data collection and sharing.[60] These trade-offs beg the questions: is it always the case that an interior solution exists, or would it sometimes be optimal for the state not to provide any government data at all? How is the optimal level of government data provision affected by economies of scope? We next illustrate qualitative implications of the model using a numerical example.

Discussion and numerical example In Figure 5, we show how the second-best government data provision changes as economies of scope become stronger, both for a case without $(\varphi = 0)$ and with $(\varphi = 0.1)$ citizens' privacy concerns. The model parameterization is the same as in our numerical example from Section 7.1. For the case with $\phi = 0$, we

[60] We emphasize disutility arising from government data collection *and* sharing because, in our model, data collection and data sharing perfectly covary (see Corollary 1). The intuition of the model would be very similar if data collection and data sharing (and their associated disutilities) were decoupled.

find that when α is below 0.7 then it is never optimal for the state to supply any government data. Therefore, when economies of scope are sufficiently low, the second-best BGP equilibrium would only feature the production of commercial software using private data alone, and no production of government software. As economies of scope become greater, so does the second-best government data supplied in equilibrium, because a higher level of government data provision to firms causes larger changes in the economy's growth rate which further compensate for the crowding out of resources from consumption. In particular, at our benchmark estimate of $\alpha = 0.8$ consistent with our empirical evidence, the second-best spending share is about 0.08. However, when privacy concerns are sufficiently strong — as is the case with $\varphi = 0.1$ — no provision of government data is justified even at our benchmark estimate.

8 Conclusion

In this paper, we provide the first evidence of a causal effect of government data on AI innovation, and study the trade-offs presented by states' AI procurement and data provision policies. These polices stimulate commercial innovation just as states' spending on space exploration and national defense did in the past. However, they work through a distinct mechanism — economies of scope arising from access to government data — and present distinct normative trade-offs.

Our analysis suggests two directions for future research, both broadening and deepening our understanding of states' role in data-intensive innovation. One natural next step is to examine the role of government data in data-intensive applications other than facial recognition. For instance, health data, collected and possessed by states in enormous quantities, can shape diagnoses, treatment, and the organization of the health sector. Geospatial data, again often collected and possessed by states, can be used in AI-fueled predictions that can transform sectors including transportation, mineral extraction, and energy production. We expect that the logic of economies of scope arising from government data and some of the normative trade-offs we have highlighted will apply to these sectors as well. However, the quantitative importance of government data could differ due to differences in technology, market structure, and institutional features that govern states' data collection and sharing.

A second direction for future work is to study in greater depth the specific political economy dimensions of data-intensive innovation. One naturally wonders to what extent autocrats' investments in surveillance AI are motivated by their desire to maintain political control, and to what extent such efforts are successful. One also wonders whether the

greater data collection in surveillance states or in societies with weaker privacy norms generate a comparative advantage in AI, and if so, what the implications for trade policy would be. 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.

References

Acemoglu, Daron, "Why Do New Technologies Complement Skills? Directed Technical Change and Wage Inequality," *The Quarterly Journal of Economics*, November 1998, *113* (4), 1055–1089.

_ , "Directed Technical Change," *The Review of Economic Studies*, October 2002, *69* (4), 781–809.

_ **and James A Robinson**, "Economic Backwardness in Political Perspective," *American Political Science Review*, February 2006, *100* (1), 1–17.

_ **and** _ , *Why Nations Fail* The Origins of Power, Prosperity, and Poverty, New York: Crown Business, August 2012.

_ **and Pascual Restrepo**, "The wrong kind of AI? Artificial intelligence and the future of labour demand," *Cambridge Journal of Regions, Economy and Society*, December 2019, *13* (1), 25–35.

_ **, David Cutler, Amy Finkelstein, and Joshua Linn**, "Did Medicare Induce Pharmaceutical Innovation?," *American Economic Review: Papers & Proceedings*, April 2006, *96* (2), 103–107.

_ **, Philippe Aghion, Leonardo Bursztyn, and David Hemous**, "The Environment and Directed Technical Change," *American Economic Review*, February 2012, *102* (1), 131–166.

Adhvaryu, Achyuta, Anant Nyshadham, and Jorge A Tamayo, "Managerial Quality and Productivity Dynamics," *Working Paper*, March 2019, pp. 1–75.

Aghion, Philippe, Antoine Dechezleprêtre, David Hemous, Ralf Martin, and John Van Reenen, "Carbon Taxes, Path Dependency, and Directed Technical Change: Evidence from the Auto Industry," *Journal of Political Economy*, February 2016, *124* (1), 1–51.

_ **, Benjamin F Jones, and Charles I Jones**, "Artificial Intelligence and Economic Growth," *NBER Working Paper*, October 2017, pp. 1–57.

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

_ **,** _ **, and** _ **, eds**, *The Economics of Artificial Intelligence An Agenda*, University of Chicago Press, 2019.

Alic, John A, Lewis M Branscomb, Harvey Brooks, and Ashton B Carter, *Beyond Spinoff* Military and Commercial Technologies in a Changing World, Harvard Business Press, 1992.

Azoulay, Pierre, Erica Fuchs, Anna Goldstein, and Michael Kearney, "Funding Breakthrough Research: Promises and Challenges of the "ARPA Model"," *NBER Working Paper*, June 2018, pp. 1–32.

_ , **Joshua S Graff Zivin, Danielle Li, and Bhaven N Sampat**, "Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules," *The Review of Economic Studies*, June 2018, *86* (1), 117–152.

Bai, Chong-En, Chang-Tai Hsieh, and Zheng Song, "Special Deals with Chinese Characteristics," *Working Paper*, May 2019, pp. 1–48.

Barro, Robert J, "Government Spending in a Simple Model of Endogeneous Growth," *Journal of Political Economy*, October 1990, *98* (5), 1–24.

Bartelme, Dominick, Arnaud Costinot, Dave Donaldson, and Andres Rodriguez-Clare, "The Textbook Case for Industrial Policy: Theory Meets Data," *Working Paper*, August 2019, pp. 1–69.

Barwick, Panle Jia, Myrto Kalouptsidi, and Nahim Bin Zahur, "China's Industrial Policy: an Empirical Evaluation," *Working Paper*, July 2019, pp. 1–68.

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb, "Are Ideas Getting Harder to Find?," *American Economic Review*, April 2020, *110* (4), 1104–1144.

_ , **John Van Reenen, and Heidi L Williams**, "A Toolkit of Policies to Promote Innovation," *Journal of Economic Perspectives*, August 2019, *33* (3), 163–184.

Bombardini, Matilde, Bingjing Li, and Ruoying Wang, "Import Competition and Innovation: Evidence from China," *Working Paper*, January 2018, pp. 1–44.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," *Working Paper*, 2017, pp. 1–35.

Brandt, Loren and Thomas G Rawski, *China's Great Economic Transformation*, Cambridge University Press, April 2008.

Chen, Yuyu, Avraham Ebenstein, Michael Greenstone, and Hongbin Li, "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*, 2013.

Cheng, Hong, Ruixue Jia, Dandan Li, and Hongbin Li, "The Rise of Robots in China," *Journal of Economic Perspectives*, May 2019, *33* (2), 71–88.

Clemens, Jeffrey and Parker Rogers, "Demand Shocks, Procurement Policies, and the Nature of Medical Innovation: Evidence from Wartime Prosthetic Device Patents," *NBER Working Paper*, January 2020, pp. 1–94.

Costinot, Arnaud, Dave Donaldson, Margaret Kyle, and Heidi L Williams, "The More We Die, The More We Sell? A Simple Test of the Home-Market Effect," *The Quarterly Journal of Economics*, January 2019, *134* (2), 843–894.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv.org*, October 2018.

Farboodi, Maryam, Toxana Mihet, Thomas Philppon, and Laura L Veldkamp, "Big Data and Firm Dynamics," *NBER Working Paper*, January 2019, pp. 1–11.

Fisman, Raymond and Yongxiang Wang, "The Mortality Cost of Political Connections," *The Review of Economic Studies*, September 2015, *82* (4), 1346–1382.

Giorcelli, Michela, "The Long-Term Effects of Management and Technology Transfers," *American Economic Review*, January 2019, *109* (1), 121–152.

Goldfarb, Avi and Daniel Trefler, "Artificial intelligence and international trade," in "The Economics of Artificial Intelligence: An Agenda," University of Chicago Press, 2018, pp. 463–492.

Greenstein, Shane, *How the Internet Became Commercial* Innovation, Privatization, and the Birth of a New Network, Princeton University Press, October 2015.

Gross, Daniel P and Bhaven N Sampat, "Inventing the Endless Frontier: the Effects of the World War Ii Research Effort on Post-War Innovation," *NBER Working Paper*, June 2020, pp. 1–58.

Guriev, Sergei and Daniel Treisman, "Informational Autocrats," *Journal of Economic Perspectives*, November 2019, *33* (4), 100–127.

Hanlon, W Walker, "Necessity Is the Mother of Invention: Input Supplies and Directed Technical Change," *Econometrica*, February 2015, *83* (1), 67–100.

_ , "The Persistent Effect of Temporary Input Cost Advantages in Shipbuilding, 1850-1911," *Journal of the European Economic Association*, 2020, pp. 1–86.

He, Guojun, Shaoda Wang, and Bing Zhang, "Watering Down Environmental Regulation in China," *The Quarterly Journal of Economics*, June 2020, *135* (4), 2135–2185.

Hemous, David, "The dynamic impact of unilateral environmental policies," *Journal of International Economics*, November 2016, *103* (C), 80–95.

Howell, Sabrina T, "Financing Innovation: Evidence from R&D Grants," *American Economic Review*, April 2017, *107* (4), 1136–1164.

Jia, Ruixue, Masayuki Kudamatsu, and David Seim, "Political Selection in China: the Complementary Roles of Connections and Performance," *Journal of the European Economic Association*, April 2015, *13* (4), 631–668.

Jones, Charles I and Christopher Tonetti, "Nonrivalry and the Economics of Data," *Working Paper*, October 2018, pp. 1–43.

Juhàsz, Réka, "Temporary Protection and Technology Adoption: Evidence from the Napoleonic Blockade," *American Economic Review*, November 2018, *108* (11), 3339–3376.

Kalouptsidi, Myrto, "Detection and Impact of Industrial Subsidies: The Case of Chinese Shipbuilding," *The Review of Economic Studies*, August 2017, *85* (2), 1111–1158.

Khandelwal, Amit K, Peter K Schott, and Shang-Jin Wei, "Trade Liberalization and Embedded Institutional Reform: Evidence from Chinese Exporters," *American Economic Review*, October 2013, *103* (6), 2169–2195.

Lane, Nathaniel, "Manufacturing Revolutions: Industrial Policy and Networks in South Korea," *Working Paper*, January 2017, pp. 1–90.

_ , "The New Empirics of Industrial Policy," *Journal of Industry, Competition and Trade*, January 2020, *59* (2), 1–26.

Lau, Lawrence J, Yingyi Qian, and Gerard Roland, "Reform without Losers: An Interpretation of China's Dual-Track Approach to Transition," *Journal of Political Economy*, February 2000, *108* (1), 120–143.

Li, Hongbin and Li-An Zhou, "Political turnover and economic performance: the incentive role of personnel control in China," *Journal of Public Economics*, September 2005, *89* (9-10), 1743–1762.

Li, Weijia, "Rotation, Performance Rewards, and Property Rights," *Working Paper*, February 2019, pp. 1–75.

Liu, Ernest, "Industrial Policies in Production Networks," *The Quarterly Journal of Economics*, August 2019, *134* (4), 1883–1948.

Mitrunen, Matti, "War Reparations, Structural Change, and Intergenerational Mobility," *Working Paper*, January 2019, pp. 1–59.

Moretti, Enrico, Claudia Steinwender, and John Van Reenen, "The Intellectual Spoils of War? Defense R&D, Productivity and International Spillovers," *NBER Working Paper*, November 2019, pp. 1–76.

Moser, Petra, "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs," *American Economic Review*, August 2005, *95* (4), 1214–1236.

Murphy, Kevin M, Andrei Shleifer, and Robert W Vishny, "Industrialization and the Big Push," *Journal of Political Economy*, October 1989, *97* (5), 1–25.

Nagaraj, Abhishek and Scott Stern, "The Economics of Maps," *Journal of Economic Perspectives*, February 2020, *34* (1), 196–221.

Nagle, Frank, "Government Technology Policy, Social Value, and National Competitiveness," *Working Paper*, March 2019, pp. 1–52.

North, Douglass C, John Joseph Wallis, and Barry R Weingast, *Violence and Social Orders* A Conceptual Framework for Interpreting Recorded Human History, Cambridge: Cambridge University Press, February 2009.

Panzar, John C and Robert D Willig, "Economies of Scope," *American Economic Review: Papers & Proceedings*, May 1981, *71* (2), 1–6.

Perrault, Raymond, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles, "The AI Index 2019 Annual Report," Technical Report, AI Index Steering Committee, Human-Centered AI Institute, Stanford University December 2019.

Popp, David, "Induced Innovation and Energy Prices," *American Economic Review*, February 2002, *92* (1), 160–180.

Roberts, Mark J, Daniel Yi Xu, Xiaoyan Fan, and Shengxing Zhang, "The Role of Firm Factors in Demand, Cost, and Export Market Selection for Chinese Footwear Producers," *The Review of Economic Studies*, November 2017, *85* (4), 2429–2461.

Rodrik, Dani, "Industrial Development: Stylized Facts and Policies," *Working Paper*, August 2007, pp. 1–33.

Scott, James C, *Seeing Like a State* How Certain Schemes to Improve the Human Condition Have Failed, Yale University Press, 1998.

Shleifer, Andrei and Robert W Vishny, *The Grabbing Hand* Government Pathologies and Their Cures, Harvard University Press, 2002.

Slavtchev, Viktor and Simon Wiederhold, "Does the Technological Content of Government Demand Matter for Private R&D? Evidence from US States," *American Economic Journal: Macroeconomics*, April 2016, *8* (2), 45–84.

Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti, "Growing Like China," *American Economic Review*, February 2011, *101* (1), 196–233.

Tirole, Jean, "Digital Dystopia," *Working Paper*, February 2020, pp. 1–36.

Tsai, Lily L, *Accountability without Democracy* Solidary Groups and Public Goods Provision in Rural China, Cambridge University Press, August 2007.

Wei, Shang-Jin, Zhuan Xie, and Xiaobo Zhang, "From "Made in China" to "Innovated in China": Necessity, Prospect, and Challenges," *Journal of Economic Perspectives*, February 2017, *31* (1), 49–70.

Williams, Heidi L, "Intellectual Property Rights and Innovation: Evidence from the Human Genome," *Journal of Political Economy*, February 2013, *121* (1), 1–27.

Zuboff, Shoshana, *The Age of Surveillance Capitalism* The Fight for a Human Future at the New Frontier of Power, PublicAffairs, January 2019.

ONLINE APPENDIX

Figure A.1: Example of AI firm record from *Tianyancha* (excerpt).

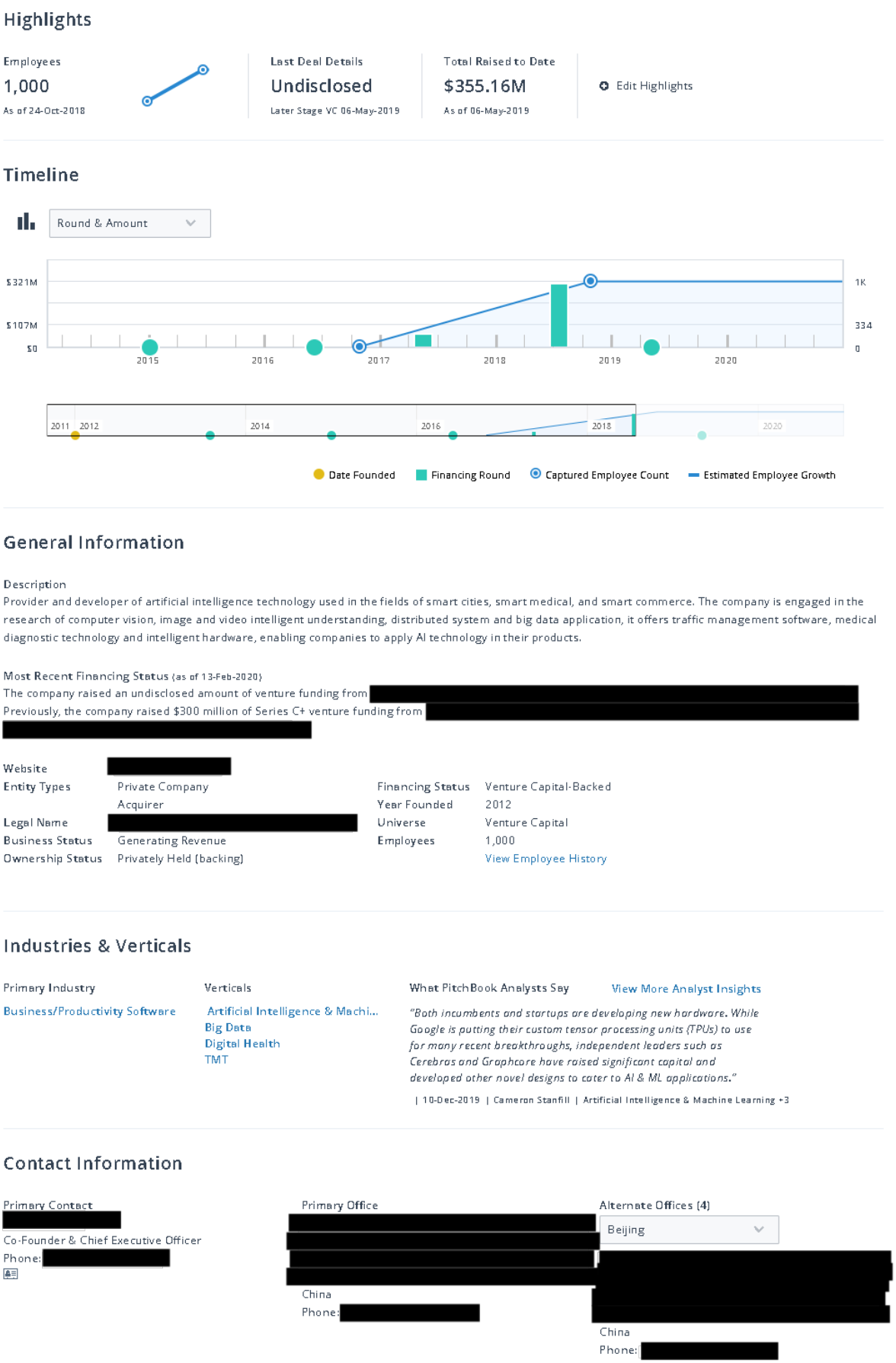


Figure A.2: Example of AI firm record from *Pitchbook* (excerpt).

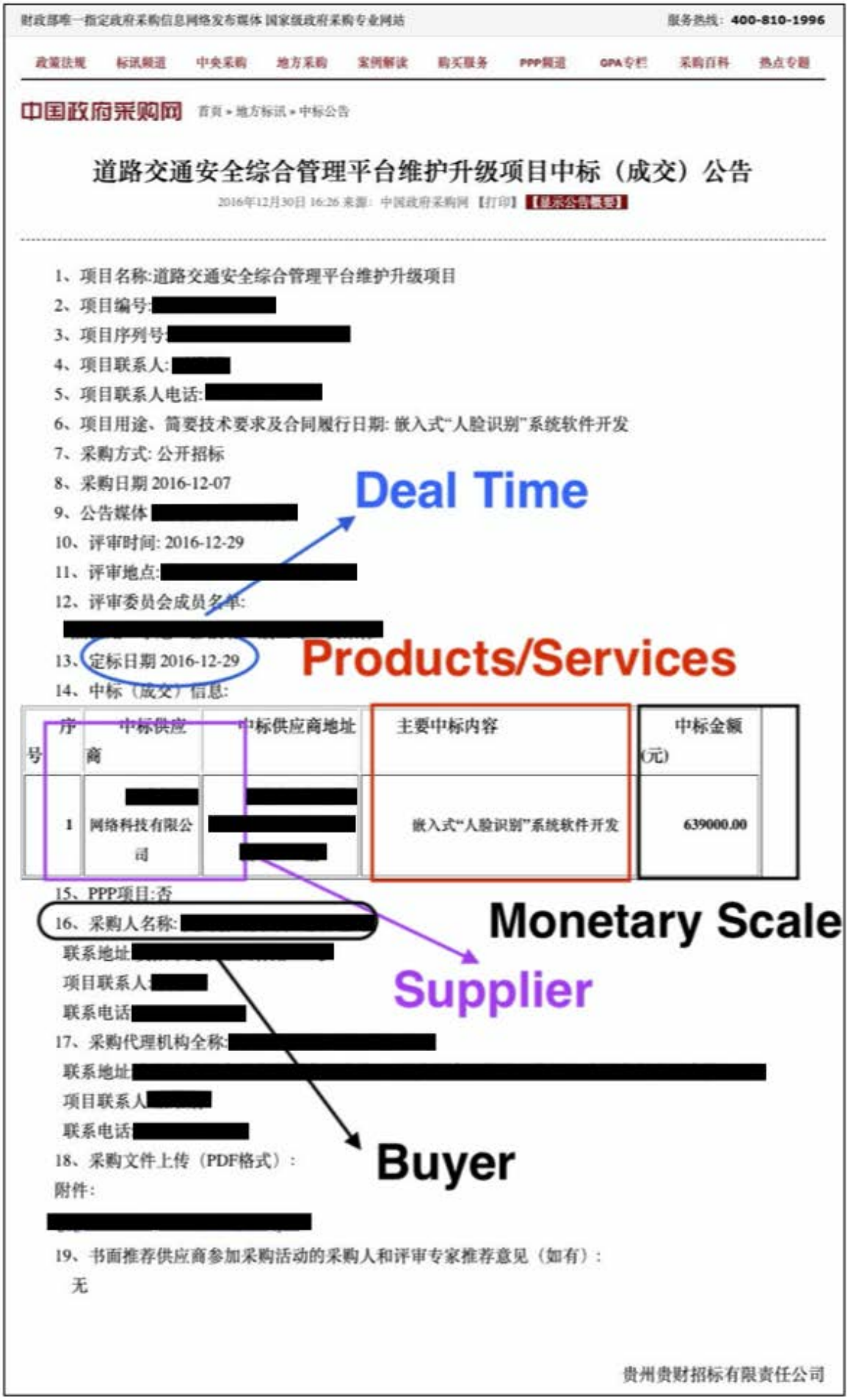


Figure A.3: Example of a procurement contract record; source: Chinese Government Procurement Database.

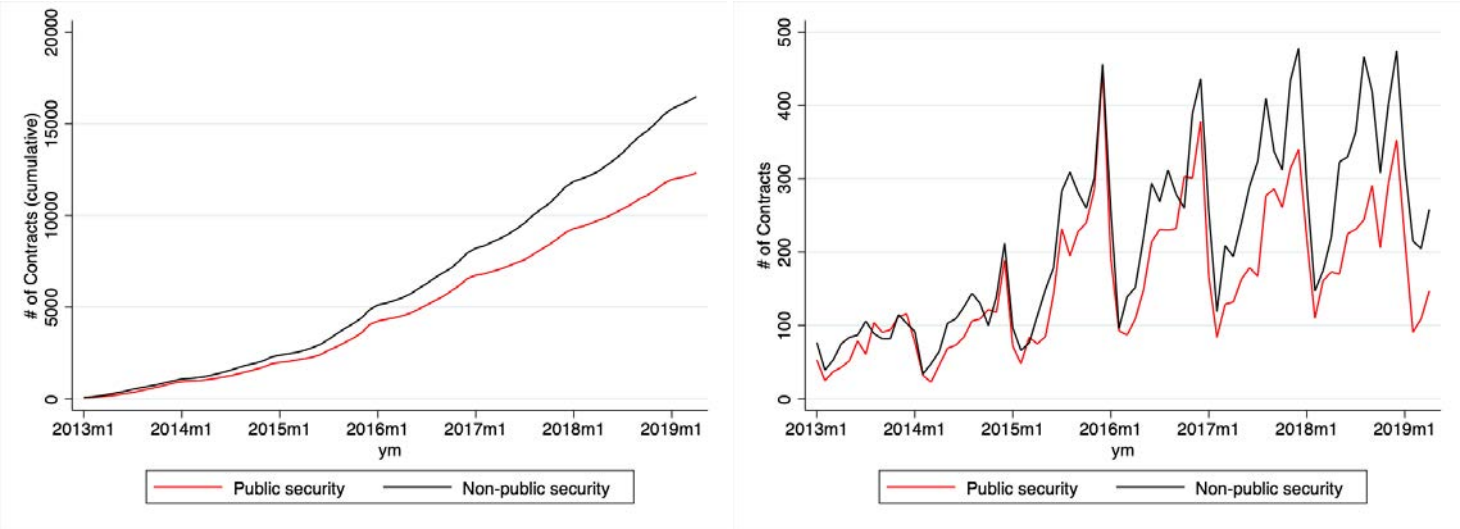


Figure A.4: Cumulative number of public security and non-public security contracts (left panel), and the flow of new contracts signed in each month (right panel).

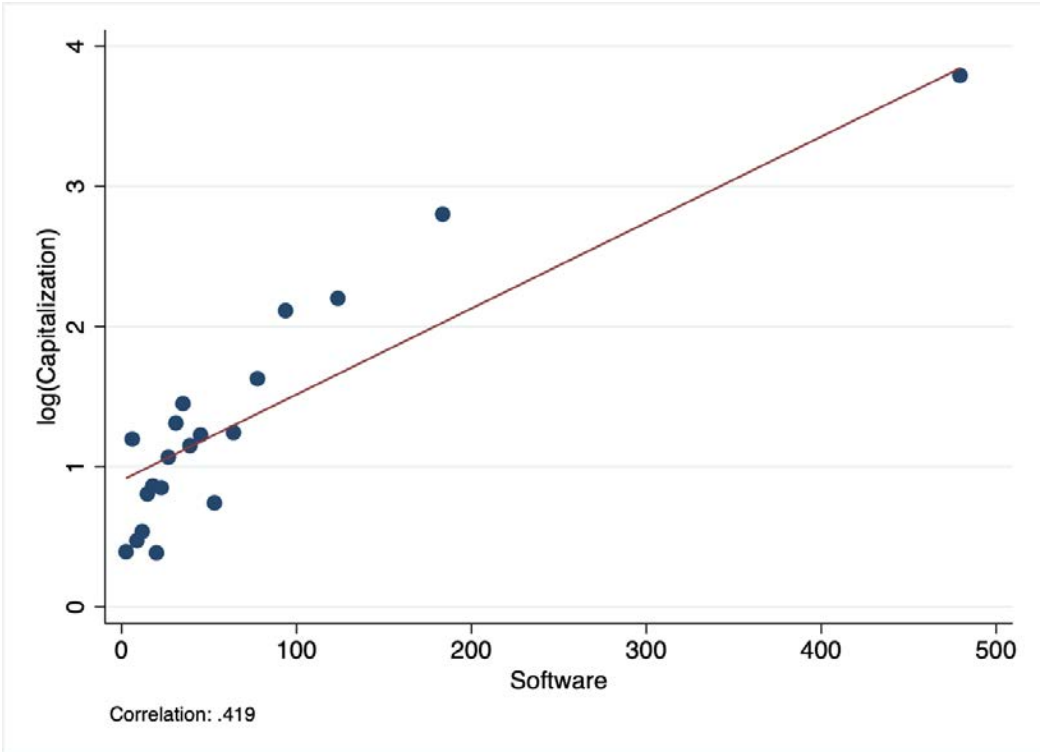
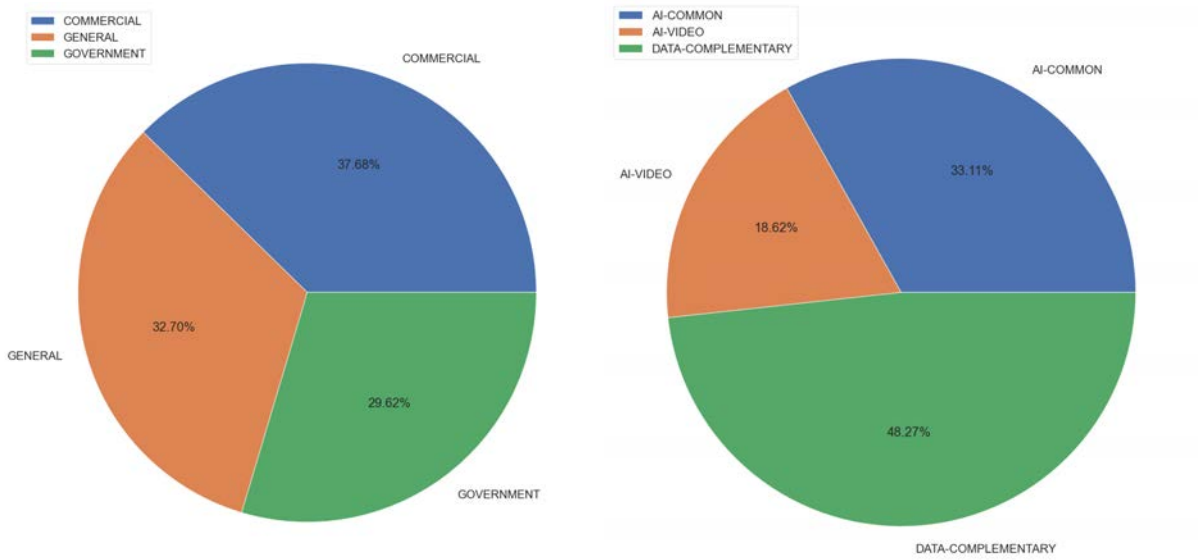


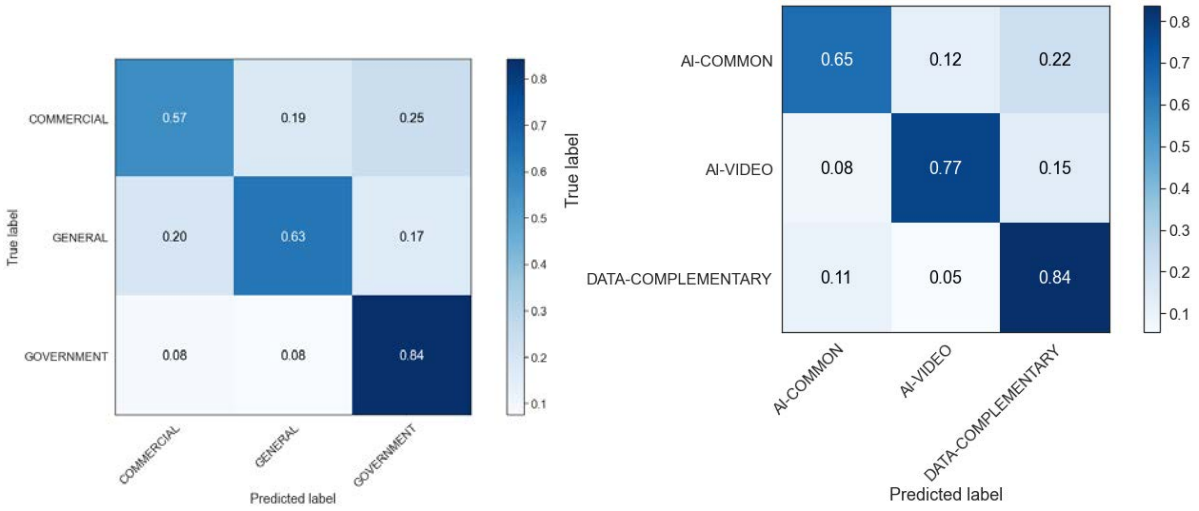
Figure A.5: Binscatter plot at the firm level of log(firm capitalization) and amount of software produced.



(a) Customers

(b) Function

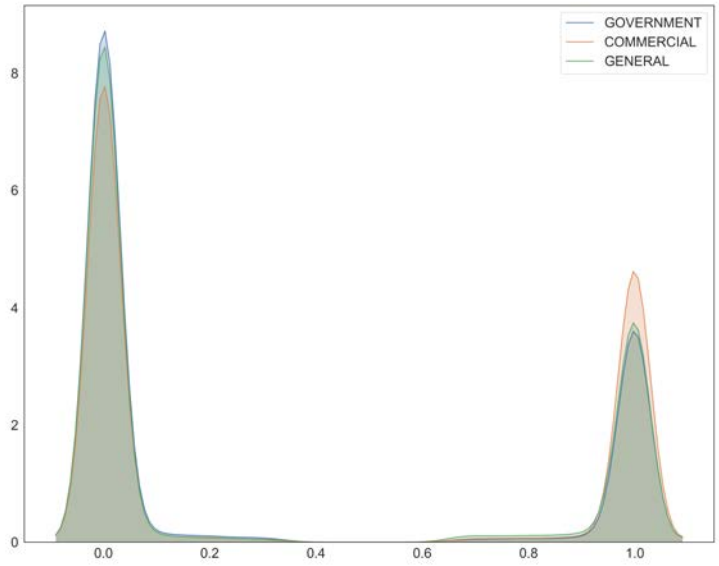
Figure A.6: Summary statistics of categorization outcomes for software categorizations based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization by customers; bottom panel shows categorization by function.



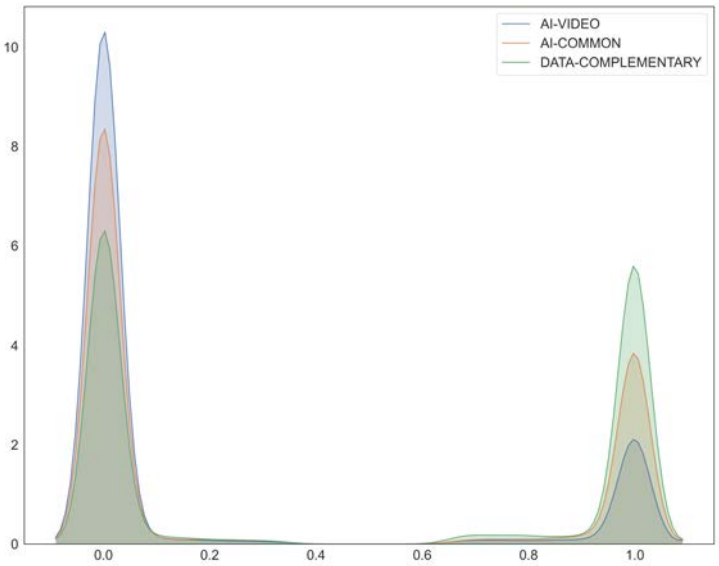
(a) Customers

(b) Functions

Figure A.7: Confusion matrix of categorization outcomes for software categorizations. True labels are based on training set constructed by human categozations (performed by two individuals). Predicted labels are outputs based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization by customers; bottom panel shows categorization by function.



(a) Customers



(b) Function

Figure A.8: Probability density plots of software categorizations based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization by customers; bottom panel shows categorization by function.

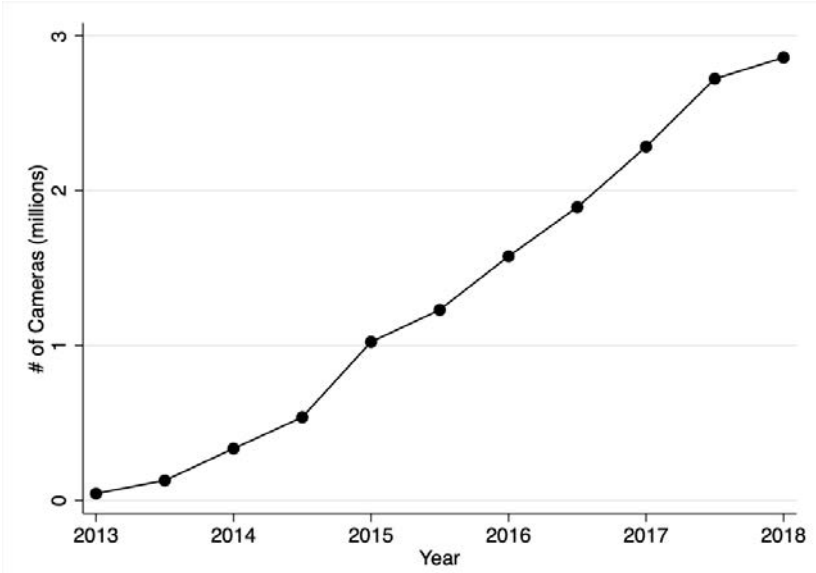
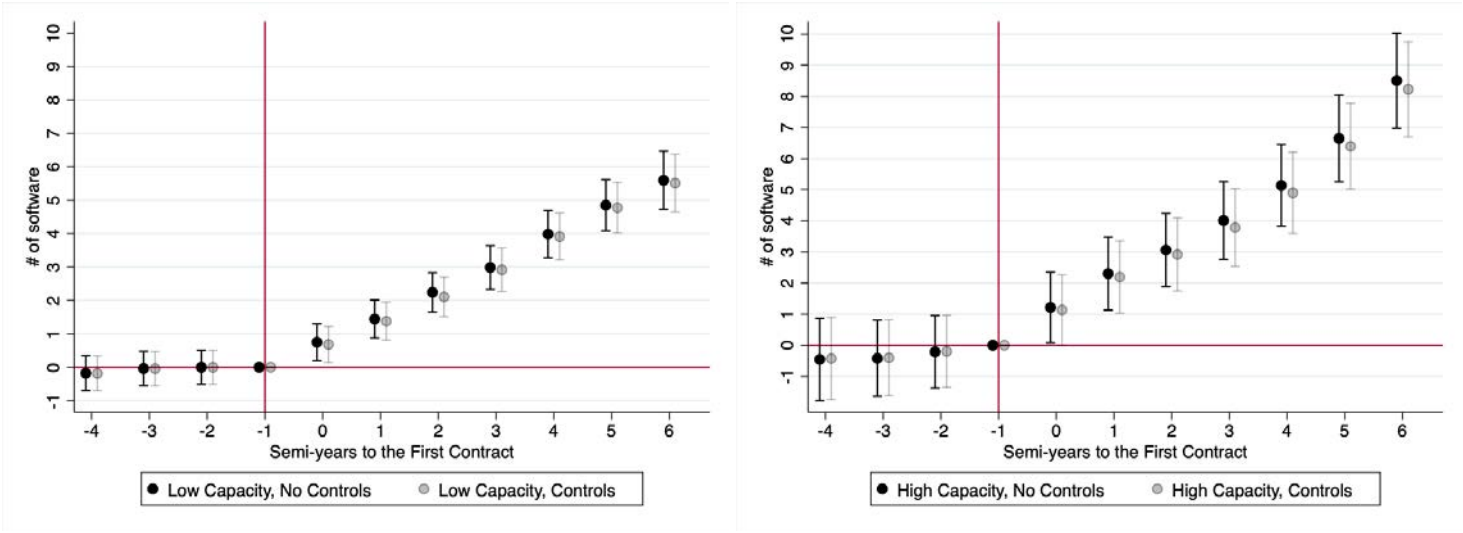
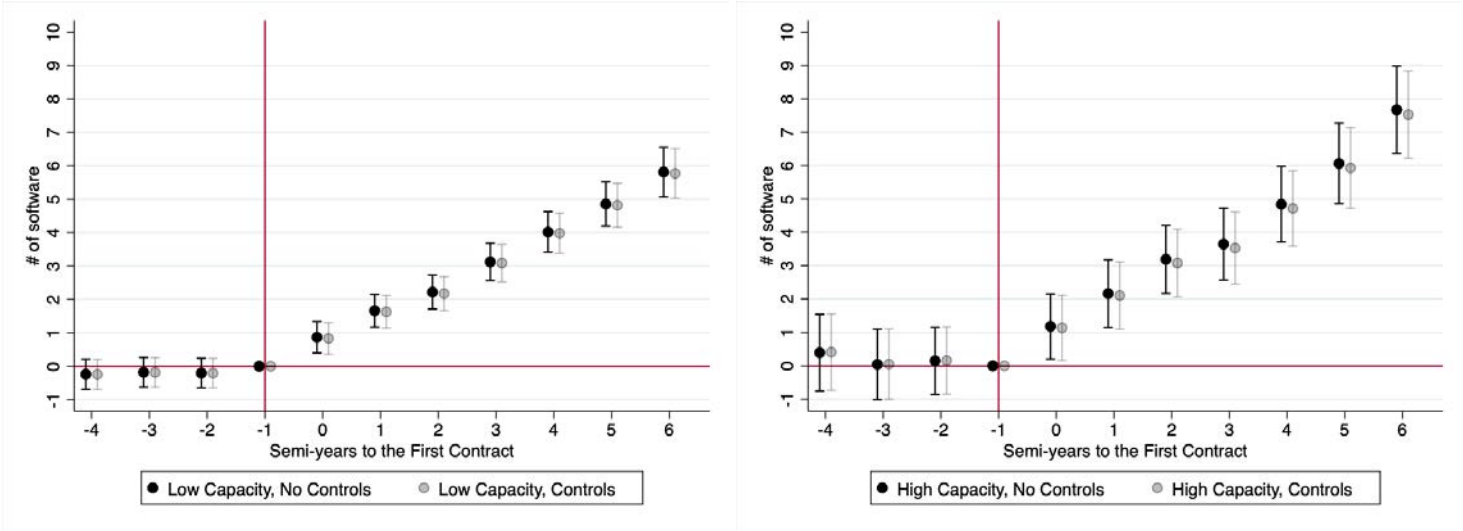


Figure A.9: Number of new public surveillance cameras in China since 2013, as measured by government procurement contracts for cameras. Source: Chinese Government Procurement Database.

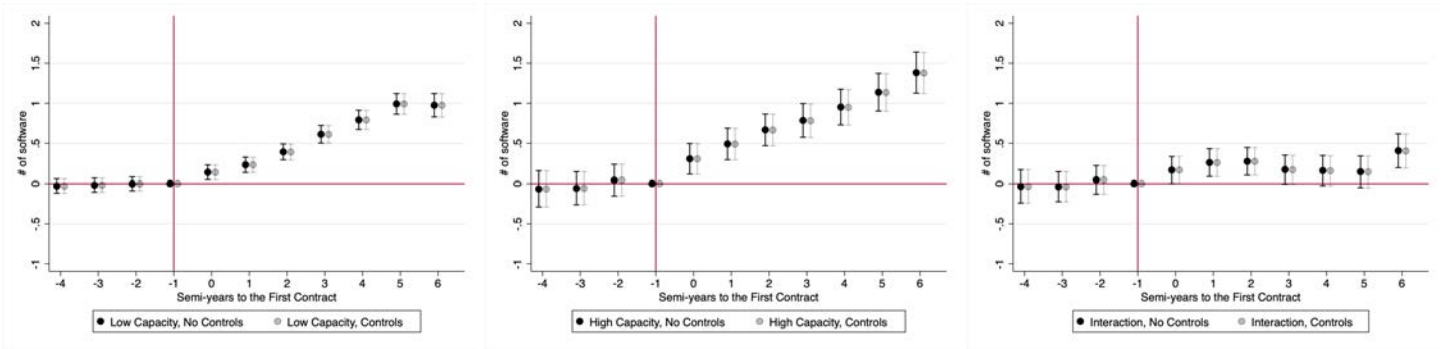


(a) Government



(b) Commercial

Figure A.10: Software development intended for government (Panel A) or for commercial uses (Panel B), resulting from data-rich public security contracts (right column) and data-scarce public security contracts (left column), controlling for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.



Panel B: Facial recognition AI involving video, high capacity vs. low capacity public security contracts

Figure A.11: Facial recognition software development that involves video (N-to-N matching). Results are presented for public security contracts that are data-scarce (left column), data-rich (middle column), and the difference (right column). All figures control for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.

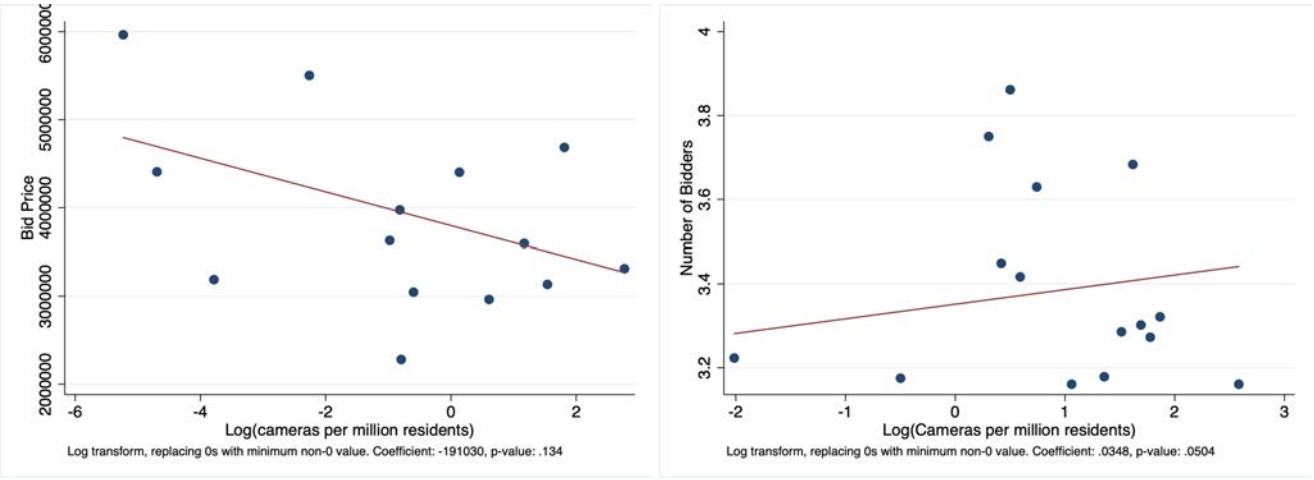
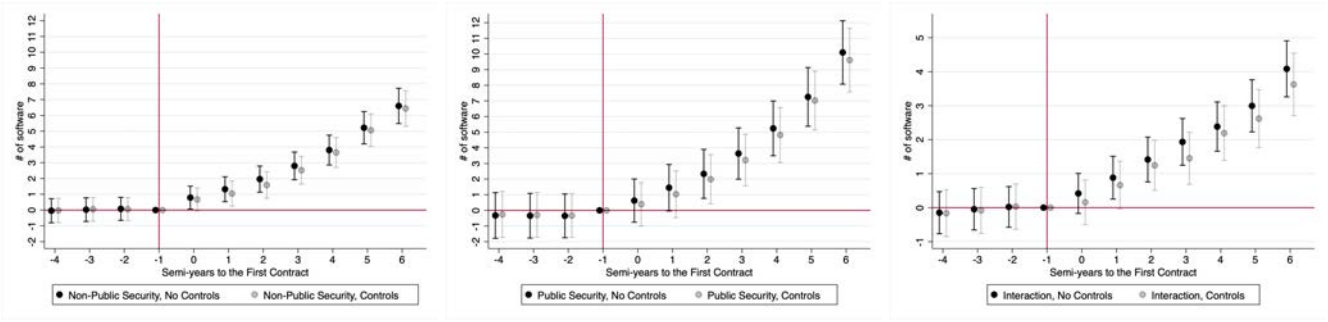
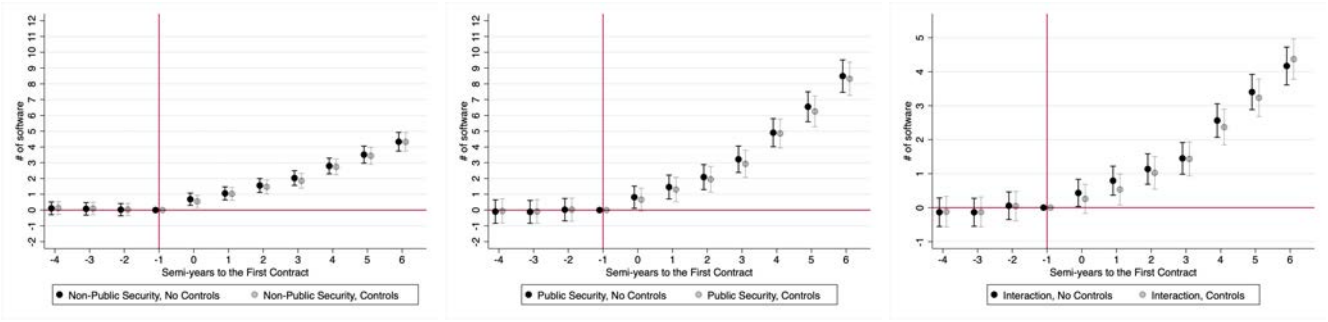


Figure A.12: Binned scatterplots of size of bid versus prefecture surveillance capacity, conditional on company fixed effects (left); and of number of bidders versus prefecture surveillance capacity (right).

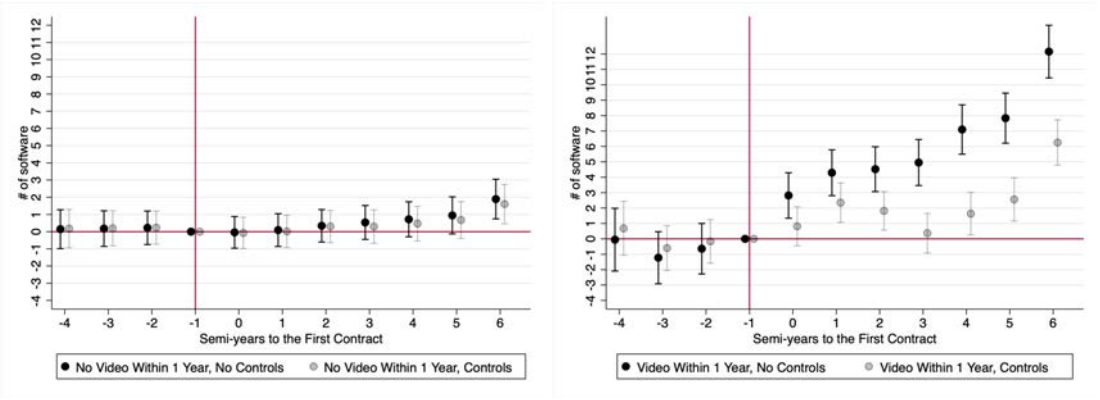


(a) Government



(b) Commercial

Figure A.13: Software development intended for government (Panel A) or for commercial uses (Panel B), resulting from public security contracts (right column), non-public security contracts (middle column), and the interaction (right column) controlling for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.



Data-complementary, split by government AI video production in year 1

Figure A.14: Data-complementary software production resulting from public security contracts that led to government video facial recognition AI software within 1 year (right column), and public security contracts that did not (left column), controlling for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.

(a) Data-complementary (b) Government (c) Commercial

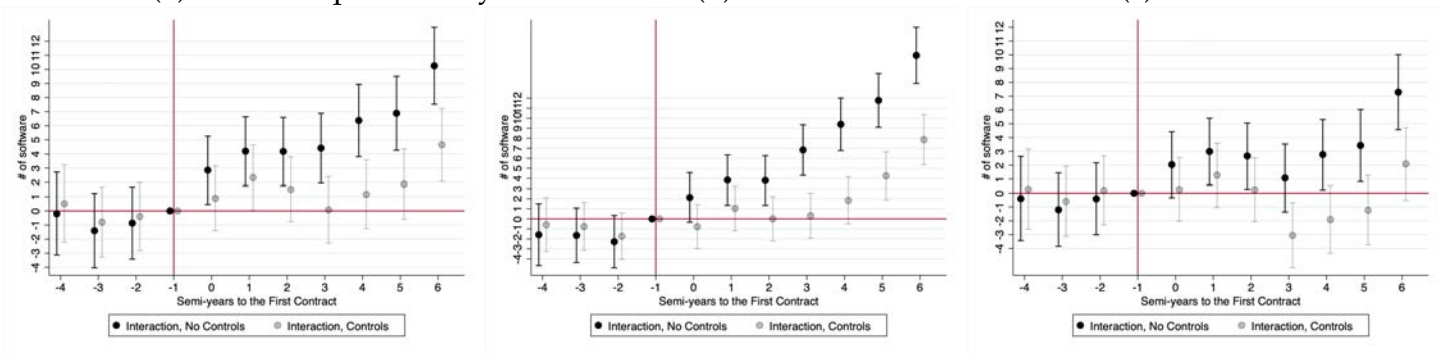


Figure A.15: Differential data-complementary software (left column) and differential AI software development intended for government (middle column) or for commercial uses (right column), resulting from public security contracts that led to government video facial recognition AI software within 1 year, relative to public security contracts that did not, controlling for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.

Table A.1: List of core variables

English name	Chinese name	Source
Panel A: Raw data		
Software	软件	Chinese Ministry of Industry and Information Technology
AI firms	人工智能公司	Tianyancha, Pitchbook
Prefecture GDP	县GDP	Global Economic Data, Indicators, Charts & Forecasts (CEIC)
Prefecture population	县人口	Global Economic Data, Indicators, Charts & Forecasts (CEIC)
Fim capitalization	公司资本	Tianyancha
Firm rounds of investment funding	公司几轮投资资金	Tianyancha
Monetary size of contracts	合约金额	Chinese Government Procurement Database
Mother firm	母公司	Tianyancha
Panel B: Constructed data		
Software customer and function	软件客户和功能	Software text
Public security contracts	公安合约	Contract text
Camera capacity	摄像机容量	Contract text
Contract runner-up bidders	合约亚军	Contract text

Table A.2: Top predicted words from LSTM model — non-binary categorization of software

Panel A: Customer type

Government			Commercial			General		
Chinese	English	Freq. (%)	Chinese	English	Freq. (%)	Chinese	English	Freq. (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
交通	Traffic	.603	手机	Mobile Phone	.821	视觉	Vision	.474
威视	Prestige	.382	APP	App	.645	学习	Learning	.378
海康	Haikang	.369	IOS	IOS	.438	腾讯	Tencent	.340
平安	Safety	.351	iOS	iOS	.430	三维	3D	.312
海信	Hisense	.318	企业	Enterprise	.331	识别系统	Recognition System	.301
城市	City	.311	金蝶	Kingdee	.327	算法	Algorithm	.270
金融	Finance	.296	电子	Electronics	.307	计算	Computing	.252
安防	Safety	.281	健康	Health	.212	深度	Depth	.225
数字	Numbers	.272	自助	Self-Help	.209	无人机	Drone	.212
中心	Center	.269	手机游戏	Mobile Game	.201	实时	Real-time	.209
公交	Public Transport	.216	助手	Assistance	.196	认证	Certification	.207
社区	Community	.207	支付	Pay	.191	处理	Processing	.196
调度	Scheduling	.200	后台	Backstage	.189	引擎	Engine	.194
中控	Central Control	.191	门禁	Access Control	.176	技术	Technique	.187
人像	Portrait	.163	人工智能	AI	.174	分布式	Distributed	.183
指挥	Command	.161	车载	Vehicle	.174	仿真	Simulation	.179
辅助	Auxilary	.159	智能家居	Smart Appliance	.169	网易	Netease	.173
摄像机	Camera	.158	工业	Industry	.169	工具软件	Tool Software	.172
万达	Wanda	.148	DHC	DHC	.168	程序	Program	.170
高速公路	Highway	.148	营销	Marketing	.161	互动	Interactive	.166

Panel B: Function type

AI-Common			Data-Complementary			AI-Video		
Chinese	English	Freq. (%)	Chinese	English	Freq. (%)	Chinese	English	Freq. (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
指纹	Fingerprint	.342	存储	Storage	.206	人脸	Face	1.104
训练	Training	.203	可视化	Visualization	.167	深度	Depth	.321
管家	Housekeeper	.201	一体化	Integration	.164	抓拍	Snapshot	.310
文本	Text	.151	分布式	Distributed	.162	商汤	SenseTime	.287
高速公路	Highway	.150	仿真	Simulation	.157	考勤	Attendance	.258
虹膜	Iris	.147	医学影像	Medical Imaging	.148	科达	Kedacom	.258
汽车	Car	.143	通用	General	.144	跟踪	Track	.249
海尔	Haier	.137	集成	Integrated	.141	全景	Panoramic	.224
WPS	WPS	.134	数据管理	Data Management	.136	广电	Broadcastt	.209
翻译	Translate	.126	宇视	UTV	.136	目标	Target/Objective	.189
推荐	Recommend	.124	管控	Manage	.126	车牌	License Plate	.189
图片	Image	.119	高速	High Speed	.126	特征	Feature	.184
测量	Test	.116	媒体	Media/Medium	.125	铂亚	Platinum	.175
征信	Credit	.111	手机软件	Phone Software	.125	预警	Warning	.166
指纹识别	Fingerprint Recognition	.106	设计	Design	.117	运通	American Express	.163
作业	Operation	.106	接口	Interface	.117	指挥	Command	.158
微信	WeChat	.105	开发	Development	.116	统计	Statistics	.149
评估	Assessment	.105	服务器	Server	.116	安居	Safety	.146
灵云	AIcloud	.102	处理软件	Processing Software	.113	SDK	SDK	.141
活体	Living Body	.098	传输	Transmission	.111	布控	Deploymentt	.141

Table A.3: Summary statistics — localities with low vs. high surveillance capacities

	Low capacity localities (1)	High capacity localities (2)	Difference (3)
Panel A: Demographics			
Population (10,000 persons)	387.613	461.803	74.189
	(263.367)	(250.099)	(32.603)**
Urban population (1,000 persons)	1,434.740	1,806.922	372.183
	(1,302.286)	(1,416.332)	(171.981)**
College students (1,000 persons)	96.034	106.309	10.276
	(186.146)	(193.176)	(23.506)
College teachers (1,000 persons)	5.256	5.573	0.318
	(10.285)	(10.570)	(1.296)
Broadband household (1000s)	1,164.550	1,680.905	516.354
	(1,119.982)	(1,306.269)	(152.231)***
Mobile phone households (1000s)	4,366.004	6,113.576	1,747.572
	(4,510.161)	(5,812.991)	(617.955)***
Observations	203	102	305
Panel B: Economics			
Number of contracts	57.369	105.225	47.856
	(117.253)	(178.565)	(17.075)***
# of 1st contracts	1.719	3.010	1.291
	(4.615)	(8.179)	(0.733)*
Monetary size (10,000 RMB)	2,671.686	2,352.398	-319.288
	(9,762.651)	(9,929.068)	(1,202.745)
GDP (100 Million RMB)	1,858.525	2,991.609	1,133.085
	(2,107.872)	(3,249.163)	(320.642)***
GDP per capita (RMB)	49,138.492	68,544.117	19,405.621
	(37,714.531)	(67,582.133)	(6,261.676)***
Fiscal expenditure (million RMB)	44,718.504	56,296.723	11,578.219
	(46,643.832)	(58,102.457)	(6,295.382)*
Fiscal revenue (million RMB)	21,227.164	33,746.250	12,519.088
	(39,860.871)	(50,784.539)	(5,433.332)**
Observations	203	102	305

Notes: Localities (at city level) are divided into below (Column 1) and above (Column 2) median in terms of their province-level surveillance-related spending prior to 2015. Broadband households are households with broadband internet connections, mobile phone households are households with a mobile phone, number of 1st contracts refers to the number of firms which had their first contract in the city, while monetary size refers to the average monetary size of all contracts. Fiscal expenditure and revenue refer to spending or revenue received by the city's government.

Table A.4: Robustness – additional results

	Government	Commercial
	(1)	(2)
Panel A.1: LSTM categorization model configuration (timestep 20, vary embeddings 16, nodes 32)		
4 semiyears before	-0.268	-0.269
	(0.288)	(0.270)
6 semiyears after	6.102***	4.743***
	(0.474)	(0.444)
4 semiyears before × data-rich	-0.181	0.418
	(0.669)	(0.634)
6 semiyears after × data-rich	2.532***	2.530***
	(0.689)	(0.647)
Panel A.2: LSTM categorization model configuration (timestep 20, embeddings 32, vary nodes 16)		
4 semiyears before	-0.206	-0.353
	(0.295)	(0.310)
6 semiyears after	6.017***	4.485***
	(0.485)	(0.509)
4 semiyears before × data-rich	-0.172	0.526
	(0.685)	(0.721)
6 semiyears after × data-rich	3.190***	2.652***
	(0.706)	(0.741)
Panel B.1: LSTM categorization model threshold (70%)		
4 semiyears before	-0.133	-0.280
	(0.233)	(0.309)
6 semiyears after	3.403***	6.411***
	(0.387)	(0.507)
4 semiyears before × data-rich	-0.243	0.542
	(0.541)	(0.720)
6 semiyears after × data-rich	2.765***	2.324***
	(0.560)	(0.739)

Notes: Specifications include full set of time indicators and interactions with data-rich contracts; only selected coefficient estimates are presented. Standard errors clustered at mother firm level are reported in parentheses. Panel A varies the LSTM specification. Table 3, Columns 1-6 use the default LSTM specification with a timestep (phrase length) of 20, embedding size (number of dimensions in a vector to represent a phrase) of 32, and 32 nodes in the model. Panel A.1 presents results for the same model trained with an embedding size of 16 instead; Panel A.2 presents results for the same model trained with 16 nodes instead. The full set of combinations of results with varied model parameters do not look qualitatively different. Table 3, Columns 1-6 use the default LSTM specification with a confidence threshold for the classification of software set at 50% (e.g. the model must be at least 50% confident that a given software is government software to be classified as "government"). Panel B.1 replicates the exercise setting the threshold to be higher, at 70%. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.5: Data-rich vs. data-scarce public security contracts

	Video	Data-Complementary	Video	Data-complementary	Video	Data-complementary
	(1)	(2)	(3)	(4)	(5)	(6)
4 semiyears before	-0.030	-0.310	-0.030	-0.317	-0.032	-0.410
	(0.045)	(0.270)	(0.045)	(0.267)	(0.091)	(0.585)
3 semiyears before	-0.019	-0.118	-0.020	-0.123	-0.017	-0.241
	(0.045)	(0.266)	(0.045)	(0.262)	(0.068)	(0.435)
2 semiyears before	-0.004	-0.151	-0.004	-0.153	0.002	-0.208
	(0.044)	(0.262)	(0.044)	(0.259)	(0.048)	(0.312)
Receiving 1st contract	0.139***	0.959***	0.138***	0.853***	0.130**	1.081***
	(0.047)	(0.280)	(0.047)	(0.277)	(0.051)	(0.326)
1 semiyear after	0.232***	1.871***	0.231***	1.772***	0.207***	1.963***
	(0.049)	(0.291)	(0.049)	(0.288)	(0.070)	(0.450)
2 semiyears after	0.392***	2.576***	0.390***	2.367***	0.372***	2.689***
	(0.050)	(0.301)	(0.050)	(0.297)	(0.093)	(0.599)
3 semiyears after	0.612***	3.331***	0.611***	3.223***	0.584***	3.519***
	(0.056)	(0.336)	(0.056)	(0.331)	(0.120)	(0.768)
4 semiyears after	0.792***	4.362***	0.791***	4.248***	0.755***	4.581***
	(0.061)	(0.362)	(0.061)	(0.357)	(0.146)	(0.937)
5 semiyears after	0.992***	5.662***	0.991***	5.543***	0.945***	5.956***
	(0.066)	(0.395)	(0.066)	(0.390)	(0.173)	(1.110)
6 semiyears after	0.976***	6.383***	0.974***	6.255***	0.923***	6.676***
	(0.074)	(0.443)	(0.074)	(0.438)	(0.201)	(1.290)
4 semiyears before × data-rich	-0.036	0.130	-0.035	0.176	-0.012	0.025
	(0.105)	(0.627)	(0.105)	(0.620)	(0.095)	(0.614)
3 semiyears before × data-rich	-0.039	-0.124	-0.039	-0.099	-0.028	-0.153
	(0.095)	(0.570)	(0.095)	(0.563)	(0.086)	(0.557)
2 semiyears before × data-rich	0.046	0.118	0.046	0.136	0.048	0.082
	(0.090)	(0.540)	(0.090)	(0.534)	(0.082)	(0.528)
Receiving 1st contract × data-rich	0.168**	0.303	0.168**	0.277	0.161**	0.213
	(0.085)	(0.512)	(0.085)	(0.506)	(0.077)	(0.501)
1 semiyear after × data-rich	0.260***	0.645	0.259***	0.574	0.258***	0.582
	(0.088)	(0.528)	(0.088)	(0.521)	(0.079)	(0.515)
2 semiyears after × data-rich	0.275***	0.909*	0.275***	0.890*	0.269***	0.783
	(0.087)	(0.524)	(0.087)	(0.518)	(0.079)	(0.513)
3 semiyears after × data-rich	0.173*	0.963*	0.171*	0.711	0.178**	0.793
	(0.091)	(0.549)	(0.091)	(0.542)	(0.083)	(0.538)
4 semiyears after × data-rich	0.161*	1.256**	0.158*	0.988*	0.161*	1.090*
	(0.096)	(0.570)	(0.096)	(0.563)	(0.087)	(0.558)
5 semiyears before × data-rich	0.146	1.592***	0.143	1.303**	0.146	1.408**
	(0.100)	(0.602)	(0.100)	(0.595)	(0.091)	(0.590)
6 semiyears after × data-rich	0.407***	2.766***	0.404***	2.452***	0.397***	2.618***
	(0.108)	(0.644)	(0.108)	(0.636)	(0.098)	(0.631)
Controls	No	No	Yes	Yes	No	No
Event-study weighting	No	No	No	No	Yes	Yes

Notes: All regressions estimated on the sample of firms with first contracts with a public security agency. Baseline specification (Columns 1–2) controls for time period fixed effects and firm fixed effects. Columns 3–4 include controls for firms' pre-contract characteristics interacted with all semi-year indicators. Standard errors clustered at mother firm level are reported in parentheses. Columns 5—6 overweight (by 1000x) control groups (no contract firms) to address potential negative weighting issues in event studies (Borusyak et al., 2017). * significant at 10% ** significant at 5% *** significant at 1%.

Table A.6: Public security contracts vs. non-public security contracts

	Government	Commercial	Government	Commercial
	(1)	(2)	(3)	(4)
4 semiyears before	-0.123	0.008	-0.013	0.020
	(0.195)	(0.138)	(0.218)	(0.141)
3 semiyears before	-0.121	0.001	-0.047	0.011
	(0.191)	(0.135)	(0.214)	(0.138)
2 semiyears before	-0.108	-0.031	-0.075	-0.008
	(0.187)	(0.132)	(0.209)	(0.135)
Receiving 1st contract	0.396**	0.401***	0.246	0.348***
	(0.186)	(0.132)	(0.208)	(0.135)
1 semiyear after	0.877***	0.846***	0.655***	0.791***
	(0.200)	(0.142)	(0.223)	(0.144)
2 semiyears after	1.407***	1.355***	1.321***	1.309***
	(0.210)	(0.149)	(0.234)	(0.151)
3 semiyears after	2.146***	1.921***	2.051***	1.953***
	(0.222)	(0.158)	(0.248)	(0.160)
4 semiyears after	2.977***	2.647***	2.982***	2.627***
	(0.237)	(0.168)	(0.264)	(0.170)
5 semiyears after	3.785***	3.079***	3.894***	2.994***
	(0.256)	(0.181)	(0.284)	(0.184)
6 semiyears after	4.833***	3.728***	5.133***	3.782***
	(0.277)	(0.196)	(0.309)	(0.199)
4 semiyears before × public security	-0.145	-0.133	-0.162	-0.117
	(0.316)	(0.224)	(0.353)	(0.229)
3 semiyears before × public security	-0.051	-0.132	-0.079	-0.129
	(0.310)	(0.220)	(0.347)	(0.224)
2 semiyears before × public security	0.023	0.059	0.032	0.047
	(0.305)	(0.216)	(0.342)	(0.221)
Receiving 1st contract × public security	0.403	0.425**	0.141	0.219
	(0.300)	(0.213)	(0.335)	(0.217)
1 semiyear after × public security	0.858***	0.796***	0.781**	0.658***
	(0.320)	(0.227)	(0.358)	(0.232)
2 semiyears after × public security	1.532***	1.216***	1.143***	1.012***
	(0.336)	(0.238)	(0.375)	(0.243)
3 semiyears after × public security	1.915***	1.645***	1.462***	1.454***
	(0.350)	(0.249)	(0.391)	(0.253)
4 semiyears after × public security	2.498***	2.550***	2.101***	2.208***
	(0.369)	(0.262)	(0.413)	(0.267)
5 semiyears before × public security	3.221***	3.349***	2.573***	3.336***
	(0.390)	(0.276)	(0.436)	(0.282)
6 semiyears after × public security	4.334***	4.383***	3.588***	4.321***
	(0.419)	(0.296)	(0.469)	(0.302)
Observations	1.19e+05	1.20e+05	1.19e+05	1.20e+05

Notes: Baseline specification (Columns 1–2) controls for time period fixed effects and firm fixed effects. Columns 3-4 include controls for firms' pre-contract characteristics interacted with all semi-year indicators. Standard errors clustered at mother firm level are reported in parentheses. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.7: Evaluating alternative hypotheses

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A.1: Control for contract similarity			
4 semiyears before	-0.185	-0.219	-0.341
	(0.268)	(0.231)	(0.270)
6 semiyears after	5.667***	5.630***	6.662***
	(0.445)	(0.380)	(0.445)
4 semiyears before × data-rich	-0.267	0.603	0.178
	(0.620)	(0.539)	(0.627)
6 semiyears after × data-rich	3.213***	1.091*	3.940***
	(0.664)	(0.569)	(0.666)
Panel A.2: Control for contract size			
4 semiyears before	-0.182	-0.243	-0.317
	(0.267)	(0.231)	(0.267)
6 semiyears after	5.511***	5.769***	6.255***
	(0.441)	(0.378)	(0.438)
4 semiyears before × data-rich	-0.243	0.653	0.176
	(0.617)	(0.538)	(0.620)
6 semiyears after × data-rich	2.715***	1.759***	2.452***
	(0.638)	(0.549)	(0.636)
Panel A.3: Control for firm pre-contract size			
4 semiyears before	-0.175	-0.240	-0.310
	(0.268)	(0.231)	(0.270)
6 semiyears after	5.579***	5.824***	6.381***
	(0.444)	(0.378)	(0.443)
4 semiyears before × data-rich	-0.277	0.632	0.131
	(0.620)	(0.539)	(0.627)
6 semiyears after × data-rich	2.898***	1.871***	2.764***
	(0.642)	(0.550)	(0.644)
Panel A.4: Control for first contract's local GDP			
4 semiyears before	-0.177	-0.247	-0.312
	(0.268)	(0.231)	(0.270)
6 semiyears after	5.593***	5.927***	6.429***
	(0.444)	(0.378)	(0.443)
4 semiyears before × data-rich	-0.278	0.567	0.106
	(0.620)	(0.538)	(0.627)
6 semiyears after × data-rich	2.888***	3.054***	3.217***
	(0.650)	(0.556)	(0.652)
Panel A.5: All previous controls combined			
4 semiyears before	-0.192	-0.238	-0.358
	(0.267)	(0.230)	(0.266)
6 semiyears after	5.601***	5.767***	6.645***
	(0.443)	(0.379)	(0.439)
4 semiyears before × data-rich	-0.222	0.568	0.211
	(0.617)	(0.537)	(0.619)
6 semiyears after × data-rich	3.056***	2.320***	4.294***
	(0.672)	(0.577)	(0.669)
Panel B.1: Learning by doing - control for government pre-contract software production			
4 semiyears before	0.138	-0.076	-0.081
	(0.233)	(0.220)	(0.252)
6 semiyears after	1.769***	3.846***	3.652***
	(0.386)	(0.362)	(0.415)

4 semiyears before × data-rich	0.170	0.869*	0.489
	(0.538)	(0.514)	(0.586)
6 semiyears after × data-rich	1.477***	1.116**	1.722***
	(0.556)	(0.525)	(0.602)
Panel B.2: Learning by doing - control for same category pre-contract software production			
4 semiyears before	0.138	0.034	-0.047
	(0.233)	(0.209)	(0.253)
6 semiyears after	1.769***	2.577***	3.173***
	(0.386)	(0.344)	(0.418)
4 semiyears before × data-rich	0.170	0.841*	0.361
	(0.538)	(0.487)	(0.589)
6 semiyears after × data-rich	1.477***	1.132**	2.013***
	(0.556)	(0.498)	(0.605)
Panel B.3: Learning by doing - control for opposite category pre-contract software production			
4 semiyears before	0.080	-0.076	-0.061
	(0.250)	(0.220)	(0.256)
6 semiyears after	2.399***	3.846***	3.474***
	(0.416)	(0.362)	(0.423)
4 semiyears before × data-rich	-0.078	0.869*	0.302
	(0.579)	(0.514)	(0.596)
6 semiyears after × data-rich	2.231***	1.116**	2.111***
	(0.599)	(0.525)	(0.612)
Panel C.1: Signalling - second contract within mother firm			
4 semiyears before	-0.078	-0.431	-0.184
	(0.213)	(0.362)	(0.283)
6 semiyears after	4.606***	6.730***	6.370***
	(0.332)	(0.557)	(0.438)
4 semiyears before × data-rich	1.035	1.047	0.820
	(0.786)	(1.384)	(1.081)
6 semiyears after × data-rich	2.753***	1.975*	1.024
	(0.710)	(1.200)	(0.947)
Panel D.1: Access to commercial opportunities - drop Beijing and Shanghai			
4 semiyears before	-0.179	-0.242	-0.277
	(0.264)	(0.166)	(0.249)
6 semiyears after	5.511***	5.873***	6.286***
	(0.423)	(0.264)	(0.397)
4 semiyears before × data-rich	-0.114	0.763*	0.235
	(0.634)	(0.404)	(0.603)
6 semiyears after × data-rich	2.983***	1.118***	2.863***
	(0.641)	(0.403)	(0.605)
Panel D.2: Access to commercial opportunities - firm based outside contract province			
4 semiyears before	-0.195	-0.165	-0.293
	(0.209)	(0.245)	(0.218)
6 semiyears after	5.254***	5.862***	6.153***
	(0.333)	(0.387)	(0.346)
4 semiyears before × data-rich	-0.053	0.721	0.177
	(0.555)	(0.658)	(0.586)
6 semiyears after × data-rich	2.365***	2.747***	2.815***
	(0.542)	(0.636)	(0.567)

Notes: Specifications include full set of time indicators and interactions with data-rich contracts; only selected coefficient estimates are presented. Standard errors clustered at mother firm level are reported in parentheses. Panels B - G replicates the baseline specification in Table 3 but additionally interacts controls with time dummies, where Panel A.1 interacts contract similarity, Panel A.2 interacts the size of the contract, Panel A.3 interacts the monetary size of the firm, Panel A.4 interacts the GDP of the first contract's location, and Panel A.5 interacts with all the above controls. Panel B.1 controls for the total amount of government software produced by the firm at 1 semiyear before the contract; Panel B.2 controls for the total of amount of software indicated in the column by the firm at 1 semiyear before the contract; Panel B.3 controls for total amount of opposite category software produced by the firm at 1 semiyear before the contract, where opposite category references the other category in the pairings between government and commercial intended software, and between AI and non-AI related software. Panel C.1 restricts the sample to only subsidiary firms that did not earn the first contract within the mother firm—note that the number of observations falls to 9,300 observations in Panel C.1 from 17,400 in Table 3. Panel D.1 excludes contracts from Beijing and Shanghai (the two highest capacity prefectures/provinces), and Panel D.2 restricts the analysis to firms that have their first contract outside of their home province. * significant at 10% ** significant at 5% *** significant at 1%.

Appendix A Proofs

Appendix A.1 Existence and uniqueness of a BGP equilibrium with entry of all types of firms

Proposition 1 (Existence and Uniqueness) *Let $p_z(p_c)$ be the implicit function defined by the pricing equation* (4) *and $p_d(p_c)$ be the implicit function defined by*

$$\Pi_c(0, p_c, p_d) = \mu_z \Pi_z(p_z(p_c)). \tag{22}$$

Let $p_g(\bar{d}_g)$ be the unique solution to

$$\kappa_g \frac{\Pi_g(p_g, \bar{d}_g)}{p_g} \frac{\chi}{1+\beta(\chi-1)} = \bar{d}_g. \tag{23}$$

Given price p_c, a necessary condition for a BGP with $\tilde{N}_c/N_z > 0$ and $N_g/N_z > 0$ to exist is

$$\frac{q_c(0, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0, p_c, p_d(p_c))} < \frac{Y_c}{D_p} = \frac{\left(\frac{p_c}{1-a}\right)^{-\epsilon}}{\kappa_p} < \frac{q_c(\bar{d}_g, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g, p_c, p_d(p_c))}. \tag{24}$$

If the condition above holds, sufficient conditions for a unique equilibrium to exist are

$$\gamma > 1 + \beta(\chi - 1) \tag{25}$$

$$\Pi_g(p_g(\bar{d}_g), \bar{d}_g) + \Pi_c(\bar{d}_g, \underline{p}_c, p_d(\underline{p}_c)) - \left(2 + \frac{F}{\lambda}\right) \Pi_c(0, \underline{p}_c, p_d(\underline{p}_c)) < 0 \tag{26}$$

$$\Pi_g(p_g(\bar{d}_g), \bar{d}_g) + \Pi_c(\bar{d}_g, \bar{p}_c, p_d(\bar{p}_c)) - \left(2 + \frac{F}{\lambda}\right) \Pi_c(0, \bar{p}_c, p_d(\bar{p}_c)) > 0, \tag{27}$$

where $\underline{p}_c$ and $\bar{p}_c$ are the smallest and largest p_c such that $p_z(p_c), p_d(p_c)$ are strictly positive.

We now proceed to prove this proposition. From the representative household's Euler equation, we obtain that in a BGP:

$$r = \theta\eta + \rho \tag{28}$$

Moreover, market clearing in the goods and data markets requires:[1]

$$\tilde{N}_c q_c(0, p_c, p_d)^{1-\frac{1}{\chi}} + N_g q_c(\bar{d}_g, p_c, p_d)^{1-\frac{1}{\chi}} = Y_c = \left(\frac{p_c}{1-a}\right)^{-\epsilon} Y \tag{29}$$

$$N_z q_z(p_z)^{1-\frac{1}{\chi}} = Y_z = \left(\frac{p_z}{a}\right)^{-\epsilon} Y \tag{30}$$

$$N_g q_g(\bar{d}_g, p_g)^{1-\frac{1}{\chi}} = Y_g = G \tag{31}$$

$$\tilde{N}_c d_p(0, p_c, p_d) + N_g d_p(\bar{d}_g, p_c, p_d) = D_p = \kappa_p Y \tag{32}$$

$$N_g \bar{d}_g = D_g = \kappa_g G, \tag{33}$$

From (4), it is straightforward to see that $p_z(p_c)$ exist and has a negative derivative.

[1] Note that, as for the case of government data, we assume that private data is not sharable across firms. This can be seen from (32). Again, we abstract from the sharability of data across firms to transparently focus on the implications of the sharability of data across uses *within* a firm.

Equations (22) follows directly from the free-entry conditions of private innovators. Then, $p_d(p_c)$ exists and has a positive derivative since profit functions are increasing in their output price and decreasing in the data input price.

Equation (23) follows from the fact that $\Pi_g(p_g, \bar{d}_g) = p_g q_g(p_g, \bar{d}_g)^{1-\frac{1}{\chi}} \frac{1+\beta(\chi-1)}{\chi}$ together with market clearing in the government data and goods markets.

Then, combining the market clearing conditions in the private data and goods markets, we obtain $\tilde{N}_c/N_z$ and N_g/N_z as functions of p_c:

$$\begin{bmatrix} \frac{\tilde{N}_c}{N_z} \\ \frac{N_g}{N_z} \end{bmatrix} = \begin{bmatrix} q_c(0, p_c, p_d(p_c))^{1-\frac{1}{\chi}} & q_c(\bar{d}_g, p_c, p_d(p_c))^{1-\frac{1}{\chi}} \\ d_p(0, p_c, p_d(p_c)) & d_p(\bar{d}_g, p_c, p_d(p_c)) \end{bmatrix}^{-1} \begin{bmatrix} \frac{Y_c}{N_z} \\ \frac{D_p}{N_z} \end{bmatrix}$$
$$\begin{bmatrix} \frac{Y_c}{N_z} \\ \frac{D_p}{N_z} \end{bmatrix} = \begin{bmatrix} \left(\frac{p_c}{1-a}\right)^{-\epsilon} \\ \kappa_p \end{bmatrix} \left(\frac{p_z(p_c)}{a}\right)^{\epsilon} q_z(p_z(p_c))^{1-\frac{1}{\chi}}.$$

When the determinant of the square matrix is negative, then $\tilde{N}_c/N_z > 0$ and $N_g/N_z > 0$ if and only if the inequalities in (24) hold. We now show that the determinant is indeed negative. This requires showing that

$$\frac{q_c(\bar{d}_g, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g, p_c, p_d(p_c))} > \frac{q_c(0, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0, p_c, p_d(p_c))},$$

which is also necessary for (24) to hold.

The optimality condition for private data demand is,

$$d_p{}^{\frac{1}{\gamma}} \left(\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{1}{\gamma-1}\left(\frac{\gamma}{1+\beta(\chi-1)}-1\right)} = \frac{(1-\alpha)}{p_d} (p_c)^{\frac{\chi}{1+\beta(\chi-1)}} \beta \left(\frac{(1-\beta)}{\phi}\right)^{\frac{(1-\beta)(\chi-1)}{1+\beta(\chi-1)}}. \tag{34}$$

Then, using the definition of $q_c(.)$, we obtain

$$\begin{aligned} \frac{q_c(\bar{d}_g, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g, p_c, p_d(p_c))} &= \frac{\chi}{\chi-1} \frac{p_d(p_c)}{\beta p_c} \left(\frac{\alpha}{(1-\alpha)} \left(\frac{\bar{d}_g}{d_p(\bar{d}_g, p_c, p_d(p_c))}\right)^{\frac{\gamma-1}{\gamma}} + 1\right) && (35) \\ &= \frac{q_c(0, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0, p_c, p_d(p_c))} \left(\frac{\alpha}{(1-\alpha)} \left(\frac{\bar{d}_g}{d_p(\bar{d}_g, p_c, p_d(p_c))}\right)^{\frac{\gamma-1}{\gamma}} + 1\right) && (36) \\ &> \frac{q_c(0, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0, p_c, p_d(p_c))}. && (37) \end{aligned}$$

To conclude the proof, we need to show conditions under which p_c exists and is unique. From the free-entry conditions for software producing firms, we obtain one equation that implicitly defines p_c:

$$\Pi_g(p_g(\bar{d}_g), \bar{d}_g) + \Pi_c(\bar{d}_g, p_c, p_d(p_c)) - \left(2 + \frac{F}{\lambda}\right) \Pi_c(0, p_c, p_d(p_c)) = 0.$$

We first show that $\gamma > 1 + \beta(\chi - 1)$ is a sufficient condition for the left-hand-side (LHS) of this equation to be strictly increasing in p_c. Totally differentiating

$$\begin{aligned} \frac{\partial LHS}{\partial p_c} &= \frac{\partial \Pi_c(\bar{d}_g, p_c, p_d)}{\partial p_c} - \left(2 + \frac{F}{\lambda}\right) \frac{\partial \Pi_c(0, p_c, p_d)}{\partial p_c} + \left(\frac{\partial \Pi_c(\bar{d}_g, p_c, p_d)}{\partial p_d} - \left(2 + \frac{F}{\lambda}\right) \frac{\partial \Pi_c(0, p_c, p_d)}{\partial p_d}\right) \frac{\partial p_d}{\partial p_c} \\ &= \frac{\partial \Pi_c(\bar{d}_g, p_c, p_d)}{\partial p_c} - \left(2 + \frac{F}{\lambda}\right) \frac{\partial \Pi_c(0, p_c, p_d)}{\partial p_c} \end{aligned}$$

$$
\begin{aligned}
&+\left(\frac{\frac{\partial \Pi_c(\bar{d}_g,p_c,p_d)}{\partial p_d}}{\frac{\partial \Pi_c(0,p_c,p_d)}{\partial p_d}}-\left(2+\frac{F}{\lambda}\right)\right)\left(\mu_z\frac{\partial \Pi_z(p_z)}{\partial p_z}\frac{\partial p_z}{\partial p_c}-\frac{\partial \Pi_c(0,p_c,p_d)}{\partial p_c}\right)\\
&=q_c(\bar{d}_g,p_c,p_d)^{1-\frac{1}{\chi}}-\left(2+\frac{F}{\lambda}\right)q_c(0,p_c,p_d)^{1-\frac{1}{\chi}}\\
&+\left(\left(2+\frac{F}{\lambda}\right)-\frac{d_p(\bar{d}_g,p_c,p_d)}{d_p(0,p_c,p_d)}\right)\left(q_c(0,p_c,p_d)^{1-\frac{1}{\chi}}-\mu_z\frac{\partial \Pi_z(p_z)}{\partial p_z}\frac{\partial p_z}{\partial p_c}\right)\\
&=\left(\frac{q_c(\bar{d}_g,p_c,p_d)^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g,p_c,p_d)}-\frac{q_c(0,p_c,p_d)^{1-\frac{1}{\chi}}}{d_p(0,p_c,p_d)}\right)d_p(\bar{d}_g,p_c,p_d)\\
&-\left(\left(2+\frac{F}{\lambda}\right)-\frac{d_p(\bar{d}_g,p_c,p_d)}{d_p(0,p_c,p_d)}\right)\mu_z\frac{\partial \Pi_z(p_z)}{\partial p_z}\frac{\partial p_z}{\partial p_c}\\
&>-\left(\left(2+\frac{F}{\lambda}\right)-\frac{d_p(\bar{d}_g,p_c,p_d)}{d_p(0,p_c,p_d)}\right)\mu_z\frac{\partial \Pi_z(p_z)}{\partial p_z}\frac{\partial p_z}{\partial p_c}\\
&>0.
\end{aligned}
$$

The second equality follows from the implicit function $p_d(p_c)$, the third equality from the envelope theorem, and the fourth equality simply rearranges terms. The first inequality follows from the fact that we have shown above that $\frac{q_c(\bar{d}_g,p_c,p_d)^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g,p_c,p_d)} > \frac{q_c(0,p_c,p_d)^{1-\frac{1}{\chi}}}{d_p(0,p_c,p_d)}$. The last inequality follows from the fact that $\frac{\partial \Pi_z(p_z)}{\partial p_z}\frac{\partial p_z}{\partial p_c} < 0$ and that, from (34), we have that when $\gamma > 1+\beta(\chi-1)$, the function $d_p(\bar{d}_g,p_c,p_d)$ is weakly decreasing in $\bar{d}_g$. As such, $\frac{d_p(\bar{d}_g,p_c,p_d)}{d_p(0,p_c,p_d)} \leq 1$ and the inequality holds.

Finally, since when $\gamma > 1+\beta(\chi-1)$ the LHS is increasing in p_c, Bolzano's theorem implies that a necessary and sufficient condition for p_c to exist and be unique is that the LHS evaluated at the smallest (largest) p_c is negative (positive). The last two equations in the theorem state these conditions.

Appendix A.2 Proof of Theorem 1 and Corollary 1

Instead of showing comparative statics with respect to $p_g G/Y$ (as in Theorem 1), it is easier to first consider changes in $\bar{d}_g$ (as in Corollary 1) which result in equilibrium changes in $p_g G/Y$. Thus, we first show the comparative statics of η and n_c with respect to changes in $\bar{d}_g$. We then show that both $p_g G/Y$ and D_g/Y increase with $\bar{d}_g$. Finally, we provide intuition for the results.

Part 1. Rate of Innovation Totally differentiating the free-entry conditions, we obtain

$$
\frac{\partial p_c}{\partial \bar{d}_g}=-\frac{\frac{\partial \Pi_g(\bar{d}_g,p_g)}{\partial \bar{d}_g}+\frac{\partial \Pi_g(\bar{d}_g,p_g)}{\partial p_g}\frac{\partial p_g}{\partial \bar{d}_g}+\frac{\partial \Pi_c(\bar{d}_g,p_c,p_d)}{\partial \bar{d}_g}}{-\left(\left(2+\frac{F}{\lambda}\right)-\frac{d_p(\bar{d}_g,p_c,p_d)}{d_p(0,p_c,p_d)}\right)\mu_z\frac{\partial \Pi_z(p_z)}{\partial p_z}\frac{\partial p_z}{\partial p_c}+d_p(\bar{d}_g,p_c,p_d)\left(\frac{q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}}{d_p(\bar{d}_g,p_c,p_d)}-\frac{q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}}}{d_p(0,p_c,p_d)}\right)}
$$

$$\frac{\partial p_d}{\partial \bar{d}_g} = -\left(\mu_z \frac{\partial \Pi_z(p_z)}{\partial p_z}\frac{\partial p_z}{\partial p_c} - q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}}\right)\frac{1}{d_p(0,p_c,p_d)}\frac{\partial p_c}{\partial \bar{d}_g}.$$

We have shown in the proof of Proposition 1 that the denominator in $\frac{\partial p_c}{\partial \bar{d}_g}$ is positive. The numerator is positive as well since $p_g(\bar{d}_g)$ is increasing in $\bar{d}_g$. Taken together, they imply that

$$\frac{\partial p_z}{\partial \bar{d}_g} > 0, \frac{\partial p_d}{\partial \bar{d}_g} < 0, \frac{\partial p_c}{\partial \bar{d}_g} < 0.$$

And, finally, using the expressions for $\eta = (r-\rho)/\theta = (\mu_z \Pi_z(p_z(p_c)) - \rho)/\theta$, we get that

$$\frac{\partial \eta}{\partial \bar{d}_g} > 0.$$

Part 2. Direction of Innovation From the market clearing conditions in the commercial goods market we have

$$\left(\frac{1-a}{a}\frac{p_z}{p_c}\right)^{\epsilon} = \frac{Y_c}{Y_z} = \frac{\tilde{N}_c}{N_z}\frac{1}{q_z(p_z)^{\frac{\chi-1}{\chi}}}q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}} + \frac{N_g}{N_z}\frac{1}{q_z(p_z)^{\frac{\chi-1}{\chi}}}q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}$$
$$= \frac{N_c}{N_z}\frac{q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}} + q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}}{q_z(p_z)^{\frac{\chi-1}{\chi}}}.$$

Thus,

$$n_c = \left(\frac{1-a}{a}\frac{p_z}{p_c}\right)^{\epsilon}\frac{q_z(p_z)^{\frac{\chi-1}{\chi}}}{q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}} + q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}}$$
$$= \frac{1-a}{a}\left(\frac{1-a}{a}\frac{p_z}{p_c}\right)^{\epsilon-1}\frac{\pi_z(p_z)\frac{\chi}{1+\beta(\chi-1)}}{\pi_c(0,p_c,p_d)\chi}\frac{1}{\pi_c(\bar{d}_g,p_c,p_d)\frac{\chi}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}}}$$
$$1 + \frac{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}}{\pi_c(0,p_c,p_d)\chi}$$
$$= \frac{1-a}{a}\left(\frac{1-a}{a}\frac{p_z}{p_c}\right)^{\epsilon-1}\frac{1}{\mu_z}\frac{1}{1+\beta(\chi-1)}\frac{1}{1+(2+\frac{F}{\lambda})\frac{\pi_c(\bar{d}_g,p_c,p_d)}{\pi_g(\bar{d}_g,p_g)+\pi_c(\bar{d}_g,p_c,p_d)}\frac{1}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}}},$$

where the second line uses that $\pi_z(p_z) = p_z q_z(p_z)^{\frac{\chi-1}{\chi}}\frac{1+\beta(\chi-1)}{\chi}$, $\pi_c(0,p_c,p_d) = p_c q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}}\frac{1}{\chi}$, and $\pi_c(\bar{d}_g,p_c,p_d) = p_c q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}\frac{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}}{\chi}$. The last line follows from the free-entry conditions.

Then, differentiating

$$\frac{dlog(n_c)}{dlog(\bar{d}_g)} > (\epsilon-1)\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)} - \frac{\frac{\Pi_c(\bar{d}_g,p_c,p_d)}{\Pi_g(\bar{d}_g,p_g)+\Pi_c(\bar{d}_g,p_c,p_d)}\frac{(2+\frac{F}{\lambda})}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}}}{1+\frac{\Pi_c(\bar{d}_g,p_c,p_d)}{\Pi_g(\bar{d}_g,p_g)+\Pi_c(\bar{d}_g,p_c,p_d)}\frac{(2+\frac{F}{\lambda})}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}}}\frac{dlog\left(\frac{\Pi_c(\bar{d}_g,p_c,p_d)}{\Pi_g(\bar{d}_g,p_g)+\Pi_c(\bar{d}_g,p_c,p_d)}\right)}{dlog(\bar{d}_g)},$$

where the inequality follows from the fact that we have shown before that $\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}$ increases with $\bar{d}_g$ when $\gamma > (1+\beta(\chi-1)$, which is one of the conditions we imposed for the BGP to exist and be unique.

We have also shown before that $\frac{dlog(p_z)}{dlog(\bar{d}_g)} > 0$, $\frac{dlog(p_c)}{dlog(\bar{d}_g)} < 0$. We thus have two cases.

First, if $\frac{dlog\left(\frac{\Pi_c(\bar{d}_g,p_c,p_d)}{\Pi_g(\bar{d}_g,p_g)+\Pi_c(\bar{d}_g,p_c,p_d)}\right)}{dlog(\bar{d}_g)} > 0$, then we can directly see from the expression above that $\epsilon \geq 1$ is a sufficient condition for $\frac{dlog(n_c)}{dlog(\bar{d}_g)} > 0$.

Second, if $\frac{dlog\left(\frac{\Pi_c(\bar{d}_g,p_c,p_d)}{\Pi_g(\bar{d}_g,p_g)+\Pi_c(\bar{d}_g,p_c,p_d)}\right)}{dlog(\bar{d}_g)} < 0$, we next show that $\epsilon \geq \frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}$ is a sufficient condition for $\frac{dlog(n_c)}{dlog(\bar{d}_g)} > 0$. Since, $\frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)} > 1$, this condition is sufficient in the first case as well.

Since the term multiplying $\frac{dlog(p_z)}{dlog(\bar{d}_g)} > 0$, $\frac{dlog(p_c)}{dlog(\bar{d}_g)}$ is less than 1, we have that

$$\begin{aligned}\frac{dlog(n_c)}{dlog(\bar{d}_g)} &> (\epsilon-1)\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)} - \frac{dlog\left(\frac{\Pi_c(\bar{d}_g,p_c,p_d)}{\Pi_g(\bar{d}_g,p_g)+\Pi_c(\bar{d}_g,p_c,p_d)}\right)}{dlog(\bar{d}_g)} \\ &> (\epsilon-1)\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)} + \left(\frac{dlog\Pi_g(\bar{d}_g,p_g)}{dlog(\bar{d}_g)} - \frac{dlog(\Pi_c(\bar{d}_g,p_c,p_d))}{dlog(\bar{d}_g)}\right),\end{aligned}$$

where the last inequality follows from the fact that $\frac{\Pi_g(\bar{d}_g,p_g)}{\Pi_g(\bar{d}_g,p_g)+\Pi_c(\bar{d}_g,p_c,p_d)} < 1$.

Moreover, combining the market clearing conditions in the markets for government goods (31) and data (33), we obtain $p_g(\bar{d}_g)$ and then

$$\frac{dlog(\Pi_g(\bar{d}_g,p_g(\bar{d}_g)))}{dlog(\bar{d}_g)} = \frac{\frac{\chi}{(1-\beta)(\chi-1)}+\beta(\chi-1)}{1+\beta(\chi-1)}.$$

Furthermore, we have that

$$\begin{aligned}\frac{dlog(\Pi_c(\bar{d}_g,p_c,p_d))}{dlog(\bar{d}_g)} &= \frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}} \\ &+ \frac{p_c q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}}{\Pi_c(\bar{d}_g,p_c,p_d)}\frac{dlogp_c}{dlog\bar{d}_g} - \frac{p_d d_p(\bar{d}_g,p_c,p_d)}{\Pi_c(\bar{d}_g,p_c,p_d)}\frac{dlogp_d}{dlog\bar{d}_g} \\ &= \frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}} + \frac{dlogp_d}{dlog\bar{d}_g} \\ &+ \frac{p_c q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}}{\Pi_c(\bar{d}_g,p_c,p_d)}\left(\frac{dlogp_c}{dlogd_g} - \left(1-(1-\beta)\frac{\chi-1}{\chi}\right)\frac{dlogp_d}{dlog\bar{d}_g}\right) \\ &= \frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}} + \frac{dlogp_d}{dlog\bar{d}_g}\end{aligned}$$

$$+\frac{\chi}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}}\left(\frac{dlog p_c}{dlog\bar{d}_g}-\left(1-(1-\beta)\frac{\chi-1}{\chi}\right)\frac{dlog p_d}{dlog\bar{d}_g}\right),$$

where the first line uses the envelope theorem, the second line uses that $\Pi_c(\bar{d}_g,p_c,p_d)=p_c q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}-p_d d_p(\bar{d}_g,p_c,p_d)-\phi x(\bar{d}_g,p_c,p_d)$ and that $\phi x(\bar{d}_g,p_c,p_d)=(1-\beta)\frac{\chi-1}{\chi}p_c q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}$ and the last line uses that

$\Pi_c(\bar{d}_g,p_c,p_d)=p_c q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}\frac{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}}{\chi}$.

Also, from the free entry condition $\Pi_c(0,p_c,p_d)=\mu_z\Pi_z(p_z)$, we have that

$$\frac{dlog(p_d)}{dlog(\bar{d}_g)}=\frac{1}{\beta}\frac{dlog(p_c)}{dlog(\bar{d}_g)}-\frac{1}{\beta}\frac{1}{1+\beta(\chi-1)}\frac{dlog(p_z)}{dlog(\bar{d}_g)}. \tag{38}$$

Replacing, we obtain

$$\begin{aligned}\frac{dlog(\Pi_c(\bar{d}_g,p_c,p_d))}{dlog(\bar{d}_g)}&=\frac{\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}+\frac{\frac{dlog(p_c)}{dlog(\bar{d}_g)}+\frac{1}{\beta}\frac{dlog(p_z/p_c)}{dlog(\bar{d}_g)}}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}}\\&+\frac{dlog(p_c)}{dlog(\bar{d}_g)}\frac{(\chi-1)}{1+\beta(\chi-1)}-\frac{1}{1+\beta(\chi-1)}\frac{1}{\beta}\frac{dlog(p_z/p_c)}{dlog(\bar{d}_g)}\\&<\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}+\frac{\beta(\chi-1)}{1+\beta(\chi-1)}\frac{1}{\beta}\frac{dlog(p_z/p_c)}{dlog(\bar{d}_g)},\end{aligned}$$

where the inequality uses that $\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}+(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}<1$ and $\frac{dlog(p_c)}{dlog(\bar{d}_g)}<0$.

Finally, using the inequality on $\frac{dlog(\Pi_c(\bar{d}_g,p_c,p_d))}{dlog(\bar{d}_g)}$ and the expression for $\frac{dlog(\Pi_g(\bar{d}_g,p_g(\bar{d}_g))}{dlog(\bar{d}_g)}$,

$$\begin{aligned}\frac{dlog(n_c)}{dlog(\bar{d}_g)}&>(\epsilon-1)\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)}+\left(\frac{dlog\Pi_g(\bar{d}_g,p_g)}{dlog(\bar{d}_g)}-\frac{dlog(\Pi_c(\bar{d}_g,p_c,p_d))}{dlog(\bar{d}_g)}\right)\\&>(\epsilon-1)\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)}+\left(\frac{\frac{\chi}{(1-\beta)(\chi-1)}+\beta(\chi-1)}{1+\beta(\chi-1)}-\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}-\frac{(\chi-1)}{1+\beta(\chi-1)}\frac{dlog(p_z/p_c)}{dlog(\bar{d}_g)}\right)\\&=\left(\epsilon-\frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}\right)\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)}+\frac{\frac{\chi}{(1-\beta)(\chi-1)}+\beta(\chi-1)}{1+\beta(\chi-1)}-\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}\\&>\left(\epsilon-\frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}\right)\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)}+\frac{\frac{\chi}{(1-\beta)(\chi-1)}+\beta(\chi-1)}{1+\beta(\chi-1)}-\frac{1+\beta(\chi-1)}{\beta(\chi-1)}\beta\frac{\chi-1}{\chi}\\&=\left(\epsilon-\frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}\right)\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)}+\frac{1}{(1-\beta)(\chi-1)}+\frac{(1-\beta)(\chi-1)}{\chi},\end{aligned}$$

where the last inequality follows from the fact that $\gamma>1+\beta(\chi-1)$ is a condition for the BGP to exist and be unique. Then, to conclude, since $\frac{dlog\left(\frac{p_z}{p_c}\right)}{dlog(\bar{d}_g)}>0$, a sufficient condition for $\frac{dlog(n_c)}{dlog(\bar{d}_g)}>0$ is that $\epsilon\geq\frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}$.

Changes in $p_g G/Y$ and D_g/Y as a function of changes in $\bar{d}_g$ We now show that both government spending $p_g G/Y$ and data D_g/Y increase in a BGP whenever $\bar{d}_g$ increases.

We have that

$$\begin{aligned}\frac{G}{Y} &= \frac{1}{\kappa_g}\frac{N_G}{N_Z}\frac{\bar{d}_g}{q_z(p_z)^{1-\frac{1}{\chi}}(\frac{p_z}{a})^{\epsilon}} \\ &= \frac{1}{\kappa_g}\frac{\left(\frac{p_c}{1-a}\right)^{-\epsilon} - \kappa_p \frac{q_c(0,p_c,p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0,p_c,p_d(p_c))}}{\frac{q_c(\bar{d}_g,p_c,p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g,p_c,p_d(p_c))} - \frac{q_c(0,p_c,p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0,p_c,p_d(p_c))}}\frac{\bar{d}_g}{d_p(\bar{d}_g,p_c,p_d(p_c))} \\ &= \frac{1}{\kappa_g}\frac{1-\alpha}{\alpha}\left(\frac{\chi-1}{\chi}\beta(1-a)^{\epsilon}\frac{1}{p_d(p_c)}(p_c)^{1-\epsilon} - \kappa_p\right)\left(\frac{\bar{d}_g}{d_p(\bar{d}_g,p_c,p_d(p_c))}\right)^{\frac{1}{\gamma}},\end{aligned}$$

where the second equality follows from the solution to N_g/N_z in Theorem 1 and the last equality uses the expressions in (35).

Differentiating,

$$\begin{aligned}\frac{dlog(G/Y)}{dlog(\bar{d}_g)} &= -\frac{\frac{\chi-1}{\chi}\beta(1-a)^{\epsilon}\frac{1}{p_d(p_c)}(p_c)^{1-\epsilon}}{\frac{\chi-1}{\chi}\beta(1-a)^{\epsilon}\frac{1}{p_d(p_c)}(p_c)^{1-\epsilon} - \kappa_p}\left((\epsilon-1)\frac{dlog(p_c)}{dlog(\bar{d}_g)} + \frac{dlog(p_d)}{dlog(\bar{d}_g)}\right) \\ &\quad + \frac{1}{\gamma}\left(1 - \frac{dlog d_p(\bar{d}_g,p_c,p_d(p_c))}{dlog(\bar{d}_g)}\right) \\ &> -\left((\epsilon-1)\frac{dlog(p_c)}{dlog(\bar{d}_g)} + \frac{dlog(p_d)}{dlog(\bar{d}_g)}\right) - \frac{1}{\gamma}\frac{dlog d_p(\bar{d}_g,p_c,p_d(p_c))}{dlog(\bar{d}_g)},\end{aligned}$$

where the inequality follows from follows from $\frac{\frac{\chi-1}{\chi}\beta(1-a)^{\epsilon}\frac{1}{p_d(p_c)}(p_c)^{1-\epsilon}}{\frac{\chi-1}{\chi}\beta(1-a)^{\epsilon}\frac{1}{p_d(p_c)}(p_c)^{1-\epsilon}-\kappa_p} > 1$.

Moreover, differentiating equation (34), we obtain

$$\begin{aligned}\frac{1}{\gamma}\frac{dlog(d_p(\bar{d}_g,p_c,p_d))}{dlog(\bar{d}_g)} &= \frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + \frac{\gamma}{1+\beta(\chi-1)}(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}\left(\frac{\chi}{1+\beta(\chi-1)}\frac{dlog(p_c)}{dlog(\bar{d}_g)} - \frac{dlog(p_d)}{dlog(\bar{d}_g)}\right) \\ &\quad - \frac{1}{\gamma}\left(\frac{\gamma}{1+\beta(\chi-1)} - 1\right)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + \frac{\gamma}{1+\beta(\chi-1)}(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}.\end{aligned}$$

Replacing above and using the expression for $\frac{dlog(p_d)}{dlog(\bar{d}_g)}$ in (38), we obtain

$$\begin{aligned}\frac{dlog(G/Y)}{dlog(\bar{d}_g)} &> -\left((\epsilon + \frac{1-\beta}{\beta})\frac{dlog(p_c)}{dlog(\bar{d}_g)} - \frac{1}{\beta}\frac{1}{1+\beta(\chi-1)}\frac{dlog(p_z)}{dlog(\bar{d}_g)}\right) \\ &\quad - \frac{1}{\gamma}\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + \frac{\gamma}{1+\beta(\chi-1)}(1-\alpha)(d_p(\bar{d}_g,p_c,p_d))^{\frac{\gamma-1}{\gamma}}}\frac{1}{\beta}\frac{\frac{dlog(p_z)}{dlog(\bar{d}_g)} - (1-\beta)\frac{dlog(p_c)}{dlog(\bar{d}_g)}}{1+\beta(\chi-1)} \\ &> -\left((\epsilon + \frac{1-\beta}{\beta})\frac{dlog(p_c)}{dlog(\bar{d}_g)} - \frac{1}{\beta}\frac{1}{1+\beta(\chi-1)}\frac{dlog(p_z)}{dlog(\bar{d}_g)}\right) \\ &\quad - \frac{1}{1+\beta(\chi-1)}\frac{1}{\beta}\frac{\frac{dlog(p_z)}{dlog(\bar{d}_g)} - (1-\beta)\frac{dlog(p_c)}{dlog(\bar{d}_g)}}{1+\beta(\chi-1)}\end{aligned}$$

$$= -\left(\epsilon + \frac{1-\beta}{\beta}\left(1 - \frac{1}{(1+\beta(\chi-1))^2}\right)\right)\frac{dlog(p_c)}{dlog(\bar{d}_g)} + \frac{1}{\beta}\frac{1}{1+\beta(\chi-1)}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}\frac{dlog(p_z)}{dlog(\bar{d}_g)}$$
$$> 0,$$

where the second line follows from $\gamma > 1 + \beta(\chi - 1)$, and the last line collects terms and comes from the fact that $\frac{dlog(p_c)}{dlog(\bar{d}_g)} < 0$, $\frac{dlog(p_z)}{dlog(\bar{d}_g)} > 0$.

Finally, since $D_g/Y = \kappa_g G/Y$ and we have shown before that p_g increases with $\bar{d}_g$, the above then implies that D_g/Y and $p_g G/Y$ increase with $\bar{d}_g$.

Intuition To understand the theorem and corollary, it helps to consider the construction of a BGP equilibrium given an exogenous increase in $\bar{d}_g$ and p_g (instead of just $\bar{d}_g$ or $p_g G/Y$). The exogenous increase directly results in higher profits for those software firms obtaining government contracts through two channels. First, through higher revenues from government software production, due to both higher p_g and productivity when $\bar{d}_g$ is higher. Second, through higher revenues from private software production, due to higher productivity when government data is used.

The higher profitability results in more R&D spending in innovation. In a BGP with free entry of innovators, the opportunity cost of investment (r) has to increase until innovators are again ex-ante indifferent between introducing a new variety or not. Furthermore, the increase in r is necessary to give the signal to households to invest more of their resources, which is ultimately consistent with the BGP increase in R&D spending and, as such, in the rate of innovation η.

However, note that the above logic holds for given prices p_z, p_c, p_d. Yet, at the new higher opportunity cost r, private software only and non-software innovators would not want to introduce new varieties at the old prices. Thus, in a BGP where all three types of firms are present, it has to be that prices change such that profits increase for these other firms not directly affected by the increase in $\bar{d}_g$ and p_g. For non-software innovators, this requires that p_z increases — which then implies that p_c has to fall so that the final goods representative firm makes zero profits (equation (4)). For private software only innovators, this requires that p_d falls to compensate for both the fall in p_c and the increase in r. Finally, under the sufficient conditions for existence and uniqueness of a BGP equilibrium, η increases because the direct effect from the increase in $\bar{d}_g$ dominates the second round, general equilibrium effects of the changes in prices.

Note that the above construction determines p_c, p_z, p_d, r and η as implicit functions of $\bar{d}_g, p_g$ purely from the free-entry conditions of firms and the Euler equation for households. Next, we turn to the market clearing conditions to understand the change in the direction of private innovation n_c. From the definition of n_c together with equations (29)

and (30), we obtained before:

$$n_c = \underbrace{\left(\frac{1-a}{a}\frac{p_z}{p_c}\right)^{\epsilon}}_{=\frac{Y_c}{Y_z}} \frac{q_z(p_z)^{\frac{\chi-1}{\chi}}}{q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}} + q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}}. \tag{39}$$

Thus, there are two countervailing effects on the direction of private innovation from the increase in $\bar{d}_g, p_g$. First, the increase in p_z and decrease in p_c result in an increase in the relative demand for private software $\frac{Y_c}{Y_z}$. This demand effect biases the direction of innovation more towards private software (increases n_c). Second, the combined increase in $\bar{d}_g$ and changes in p_c, p_d may potentially result in an increase in the relative output of private software per firm (the second term decreases). This decreases n_c. The theorem shows that, if demand is sufficiently elastic ($\epsilon \geq \frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}$) and the conditions for a BGP to exist and be unique are satisfied, then the demand effect dominates and n_c increases.

To conclude the intuition for the theorem, consider the market clearing condition for government data (33). When $\bar{d}_g$ is higher, more government data needs to be supplied to those firms obtaining government contracts. Yet, at the old p_g, the increase in government software production G/Y and thus government data as a by-product $\kappa_g q_g(\bar{d}_g, p_g)^{1-\frac{1}{\chi}}$ is insufficient to match the required demand. This is because there are decreasing returns to $\bar{d}_g$ and thus the supply increases less than proportionally. Thus, it has to be that p_g increases as well so that $q_g(\bar{d}_g, p_g)^{1-\frac{1}{\chi}}$ further increases to match the required increased in $\bar{d}_g$.

Appendix B Quantitative analysis

Appendix B.1 Equilibrium conditions

Letting $i = c, g, z$, $\alpha = 1$ if $i = g$ or $i = z$, and $\bar{d}_g = 1$ if $i = z$, the profit maximization problem can be generically written as

$$\pi_i = \max_{d_p, x} \quad \frac{\chi}{\chi-1} p_i \left(\alpha (d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}} (x)^{(1-\beta)\frac{\chi-1}{\chi}} - \phi x - p_d d_p.$$

First order conditions are:

$$p_i \left(\alpha (d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}} (x)^{(1-\beta)\frac{\chi-1}{\chi}} (1-\beta) = \phi x$$

$$p_i \left(\alpha (d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}} (x)^{(1-\beta)\frac{\chi-1}{\chi}} \beta \frac{(1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}}{\alpha (d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}} = p_d d_p.$$

This implies

$$\pi_i = p_i \left(\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}} (x)^{(1-\beta)\frac{\chi-1}{\chi}} \frac{1}{\chi-1} \times \cdots \left(1+\beta(\chi-1)\frac{\alpha(d_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}}\right)$$

$$x = \left(\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}} \left(p_i\frac{1-\beta}{\phi}\right)^{\frac{1}{1-(1-\beta)\frac{\chi-1}{\chi}}}$$

$$(d_p)^{\frac{1}{\gamma}} = \frac{(1-\alpha)}{p_d}(p_i)^{\frac{1}{1-(1-\beta)\frac{\chi-1}{\chi}}} \left(\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}-1} \left(\frac{1-\beta}{\phi}\right)^{\frac{(1-\beta)\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}} \beta,$$

which then gives

$$\pi_i = (p_i)^{\frac{1}{1-(1-\beta)\frac{\chi-1}{\chi}}} \left(\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}} \left(\frac{1-\beta}{\phi}\right)^{\frac{(1-\beta)\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}} \frac{1}{\chi-1} \times \cdots \left(1+\beta(\chi-1)\frac{\alpha(d_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}}\right)$$

$$(q_i)^{\frac{\chi-1}{\chi}} = \frac{\chi}{\chi-1}\left(\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}} \left(p_i\frac{1-\beta}{\phi}\right)^{\frac{(1-\beta)\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}}$$

$$d_p = \beta\frac{\chi-1}{\chi}\frac{(1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}}{\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}}\frac{p_i}{p_d}(q_i)^{\frac{\chi-1}{\chi}}.$$

So, normalizing $\phi = (1-\beta)$, we obtain:

$$\Pi_g(\bar{d}_g, p_g) = (p_g)^{\frac{\chi}{1+\beta(\chi-1)}} (\bar{d}_g)^{\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} \frac{1+\beta(\chi-1)}{\chi-1}$$

$$Y_g = N_g\frac{\Pi_g(\bar{d}_g, p_g)}{p_g}\frac{\chi}{1+\beta(\chi-1)}$$

$$D_g = N_g\bar{d}_g$$

$$\Pi_c(\bar{d}_g, p_c, p_d) = (p_c)^{\frac{\chi}{1+\beta(\chi-1)}} \left(\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p(\bar{d}_g, p_c, p_d))^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} \frac{1}{\chi-1} \times \cdots \left(1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p(\bar{d}_g, p_c, p_d))^{\frac{\gamma-1}{\gamma}}}\right)$$

$$(d_p(\bar{d}_g, p_c, p_d))^{\frac{1}{\gamma}} = \frac{(1-\alpha)}{p_d}(p_c)^{\frac{\chi}{1+\beta(\chi-1)}} \left(\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p(\bar{d}_g, p_c, p_d))^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}-1} \beta$$

$$\Pi_c(0, p_c, p_d) = (p_c)^{\frac{\chi}{1+\beta(\chi-1)}} \left((1-\alpha)^{\frac{\gamma}{\gamma-1}} d_p(0, p_c, p_d)\right)^{\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} \frac{1}{\chi-1}$$

$$d_p(0, p_c, p_d) = \frac{1}{(p_d)^{1+\beta(\chi-1)}}(p_c)^{\chi}(1-\alpha)^{\frac{\gamma}{\gamma-1}\beta(\chi-1)}\beta^{1+\beta(\chi-1)}$$

$$Y_c = \left(N_c + \frac{1-\lambda}{\lambda}N_g\right)\frac{\chi}{\chi-1}\left((1-\alpha)(d_p(0, p_c, p_d))^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} (p_c)^{\frac{(1-\beta)(\chi-1)}{1+\beta(\chi-1)}} + N_g\frac{\chi}{\chi-1}\left(\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p(\bar{d}_g, p_c, p_d))^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} (p_c)^{\frac{(1-\beta)(\chi-1)}{1+\beta(\chi-1)}}$$

$$
\begin{aligned}
D_p &= \left(N_c + \frac{1-\lambda}{\lambda} N_g\right) d_p(0, p_c, p_d) + N_g d_p(\bar{d}_g, p_c, p_d) \\
\Pi_z(p_z) &= (p_z)^{\frac{\chi}{1+\beta(\chi-1)}} \frac{1+\beta(\chi-1)}{\chi-1} \\
Y_z &= N_z \frac{\Pi_z(p_z)}{p_z} \frac{\chi}{1+\beta(\chi-1)}.
\end{aligned}
$$

Furthermore, from the profit maximization of the final goods seller together with goods market clearing, we obtain:

$$
\begin{aligned}
Y_z &= \left(\frac{p_z}{a}\right)^{-\epsilon} Y \\
\frac{1-a}{a} \left(\frac{Y_c}{Y_z}\right)^{-\frac{1}{\epsilon}} &= \frac{p_c}{p_z} \\
\left[(1-a)^{\epsilon}(p_c)^{1-\epsilon} + a^{\epsilon}(p_z)^{1-\epsilon}\right]^{\frac{1}{1-\epsilon}} &= 1.
\end{aligned}
$$

And the remaining market clearing conditions are

$$
\begin{aligned}
G &= Y_g \\
D_g &= \kappa_g G \\
D_p &= \kappa_p Y.
\end{aligned}
$$

And the free entry conditions are

$$
\begin{aligned}
0 &= \Pi_g(\bar{d}_g, p_g) + \Pi_c(\bar{d}_g, p_c, p_d) - (2 + \frac{F}{\lambda}) \mu_z \Pi_z(p_z) \\
\Pi_c(0, p_c, p_d) &= \mu_z \Pi_z(p_z) \\
\mu_z \Pi_z(p_z) &= \theta\eta + \rho = r,
\end{aligned}
$$

where the last equality follows from the Euler equation of the representative household in a BGP.

Appendix B.2 Calibration

We externally calibrate $\theta = 2$, $\rho = 0.03$, $\chi = 6$, which are standard parameters in the literature. As for the elasticity of substitution between software and non-software intermediates, we set $\epsilon = 1$ so that the aggregate production function is Cobb-Douglas. We set $a, \mu_z, F, \kappa_g, \kappa_p$ such that the initial BGP equilibrium is symmetric: the direction of innovation is unbiased ($\frac{\tilde{N}_c}{N_z} = \frac{N_g}{N_z} = 1$) and all sectors have an identical output share ($\frac{p_c Y_c}{p_z Y_z} = \frac{p_g G}{p_z Y_z} = 1$). We assume a growth rate of 6%, which matches the annual per-capita GDP growth rate in China in recent years.

The parameters left to set are those associated with data as an input in innovation: the share of data in production β, the elasticity of substitution between government and private data γ, and the productivity of government data in private software innovation α. Admittedly, we have a large degree of uncertainty about β and γ. Our empirical evidence

on the responses of government and commercial software following the receipt of data-rich government contracts at most show that $\beta > 0$ and $\gamma < \infty$. So, for our baseline calibration, we will simply set them to $\beta = 0.8$ and $\gamma = 1 + \beta(\chi - 1) + 0.1$ which ensure that a symmetric BGP equilibrium exist.

However, given β, γ, we next show how to pin down the parameter governing economies of scope α from our empirical evidence. Fixing prices and differentiating the optimal levels of software production for those firms obtaining contracts with respect to $\bar{d}_g$, we obtain the partial equilibrium responses:

$$\Delta log(q_g) = \frac{\chi\beta}{1 + (\chi - 1)\beta} \Delta log(\bar{d}_g)$$
$$\Delta log(q_c) = \frac{\chi\beta\sigma}{1 + (\chi - 1)\beta + \gamma(1 - \sigma)} \Delta log(\bar{d}_g),$$

where

$$\sigma \equiv \frac{\alpha}{\alpha + (1 - \alpha)\frac{d_p(\bar{d}_g, p_c, p_d)}{\bar{d}_g}^{\frac{\gamma - 1}{\gamma}}}.$$

These responses are the model equivalent to those that we have estimated for high capacity contracts in Appendix Table 3, columns (1) and (2). Then, when setting the government and private data in software production in the symmetric BGP to be identical ($\bar{d}_g = d_p(\bar{d}_g, p_c, p_d)$), we obtain that $\alpha = \sigma$ and therefore:

$$\alpha = \frac{\frac{\Delta log(q_c)}{\Delta log(q_g)}}{1 - \frac{\gamma}{1 + \beta(\chi - 1) + \gamma}\left(1 - \frac{\Delta log(q_c)}{\Delta log(q_g)}\right)}.$$

We use the coefficients in Appendix Table 3, 6 Semiyears after $\times$ High-capacity, columns (1) and (2). They imply an elasticity of private to government software ($\frac{\Delta log(q_c)}{\Delta log(q_g)}$) of about 2/3. Given our parameterization, this results in $\alpha = 0.8$.