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 in September 2020
JEL No. E0,H4,L5,L63,O25,O30,O40,P00,P16,Z21

ABSTRACT

Data-intensive technologies, like AI, are increasingly widespread. We argue that the direction of innovation and growth in data-intensive economies may be crucially shaped by the state because: (i) the state is a key collector of data and (ii) data is sharable across uses within firms, potentially generating economies of scope. We study a prototypical setting: facial recognition AI in China. Collecting comprehensive data on firms and government procurement contracts, we find evidence of economies of scope arising from government data: firms awarded contracts providing access to more government data produce both more government and commercial software. We then build a directed technical change model to study the implications of government data access for the direction of innovation, growth, and welfare. We conclude with three applications showing how data-intensive innovation may be shaped by the state: both directly, by setting industrial policy; and indirectly, by choosing surveillance levels and privacy regulations.

Martin Beraja
Department of Economics
MIT
50 Memorial Drive
Cambridge, MA 02142
and NBER
martinberaja@gmail.com

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

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

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

Martin Beraja David Y. Yang Noam Yuchtman*

September 18, 2020

Abstract

Data-intensive technologies, like AI, are increasingly widespread. We argue that the direction of innovation and growth in data-intensive economies may be crucially shaped by the state because: *(i)* the state is a key collector of data and *(ii)* data is sharable across uses within firms, potentially generating economies of scope. We study a prototypical setting: facial recognition AI in China. Collecting comprehensive data on firms and government procurement contracts, we find evidence of economies of scope arising from government data: firms awarded contracts providing access to more government data produce *both* more government and commercial software. We then build a directed technical change model to study the implications of government data access for the direction of innovation, growth, and welfare. We conclude with three applications showing how data-intensive innovation may be shaped by the state: both directly, by setting industrial policy; and indirectly, by choosing surveillance levels and privacy regulations.

Keywords: data, innovation, artificial intelligence, economies of scope, directed technical change, industrial policy, China, privacy, surveillance
JEL Classification: O30, P00, E00, L5, L63, O25, O40

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

1 Introduction

Artificial intelligence and machine learning technologies ("AI" for brevity) are increasingly widespread. Because of their potential, they have attracted a great deal of attention from economists and others (see Agrawal et al., eds, 2019 for a review). Developing these technologies is *data-intensive*. The importance of data can be seen in recent breakthroughs from translation, to speech and facial recognition, to chess grand mastery: all of these were driven as much by access to massive amounts of data as by algorithmic advances.[1]

Data differs from other inputs into innovation in two important ways. First, throughout history and up to the present, states have collected massive quantities of data to fulfill their primary objectives. From administrative data that make society "legible" (Scott, 1998) and allow the state to collect taxes and provide services; to geographic and scientific data used for national defense; to surveillance data used to provide public security, among others. Tellingly, "state" is at the root of the word "statistics." Second, data can be shared across multiple uses within a firm. These two features may generate *economies of scope* from government data.[2] In particular, a firm gaining access to government data collected by the state could use that same data to develop new products for government uses as well as products intended for much larger commercial markets. In this paper, we argue that, because of these two features of data, the direction of innovation and growth in data-intensive economies may be crucially shaped by the state.

To examine the empirical relevance of the two features of data we have highlighted, we study a prototypical data-intensive sector in which the state has a significant public security interest: the facial recognition AI industry in China. We find evidence of economies of scope arising from government data: following the receipt of a government contract to supply AI software, firms produce more software both for *government* and *commercial* purposes when the contract provides access to more government data. To study the *aggregate* implications of firms' ac-

[1] See Sejnowski (2018). Kai-Fu Lee (former director of Microsoft Research Asia and president of Google China) has even argued that, as opposed to researchers, "... it is data that is crucial to the implementation of AI technologies ..." (source: `https://bit.ly/34gJkgu`).

[2] 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. Seminal work by Panzar and Willig (1981) shows how economies scope may arise when inputs are sharable.

cess to government data, we build a general equilibrium directed technical change model where some firms choose to engage in data-intensive innovation, the state and private sector demand data-intensive software to produce "surveillance" services and consumption goods, respectively, and government data gives rise to economies of scope. We show that increasing the amount of government data provided to firms can indeed increase the economy's growth rate and bias the direction of private innovation towards data-intensive software. However, because innovation crowds-out resources from consumption, government data provision increases welfare only when economies of scope are sufficiently strong. We conclude with three applications which illustrate the varied ways that data-intensive innovation may be shaped by the state: both directly, by setting industrial policies; and indirectly, by choosing public surveillance levels as well as enacting privacy regulations. These applications demonstrate that the welfare implications of government data collection and provision are further complicated by potential misalignment between citizens' and states' preferences for surveillance and privacy.

Our paper begins by presenting a simple conceptual framework where economies of scope in data-intensive innovation arise from government data being sharable across multiple uses. We derive a key prediction that guides our subsequent empirical analysis: a government contract that results in an exogenous increase in the government data available to a firm will lead to increased production of *both* government and commercial software. Yet, we note that economies of scope may not arise even when government data can be shared across uses. For instance, firms may not increase commercial software production upon receipt of a government contract if, in order to fulfill it, the firm needs to reallocate substantial resources towards government software production and away from commercial software production.

The facial recognition AI industry in China is a uniquely suited empirical context to study this question. Firms developing facial recognition software require large datasets.[3] The Chinese state both collects huge amounts of personal data and demands facial recognition software for surveillance purposes. A firm receiving a government contract would thus receive access to government data which is not

[3]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.

publicly available, using this data to develop the software it was contracted to produce. For example, when obtaining a contract with a police department to produce surveillance software, it could receive access to video from street cameras, and potentially a database of labeled personal images as well. It could then develop surveillance software by training an AI algorithm that matches individuals across video feeds or from video feeds to labeled images. Crucially, the detection of individuals from video (or photo) data is also key to any *commercial* facial recognition AI application, for instance, facial recognition platforms for retail stores. Therefore, to the extent that the government data (or fine-tuned detection 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 data challenges. First, linking AI firms to government contracts. To do so, we collect data on (approximately) the universe of Chinese facial recognition AI firms and link this data to a separate database of Chinese government contracts, issued by all levels of the government. Second, quantifying AI firms' software production and, as important, classifying firms' software by intended use. We do this by compiling data on all Chinese facial recognition AI firms' software development based on the digital product registration records maintained by the Chinese government. Using a Recurrent Neural Network model, we categorize software products based on whether they are directed towards the commercial market or government use. Third, measuring the amount of government data to which firms have access. To do this, we construct two proxies for the data-richness of an AI contract. We begin by distinguishing among government contract awarding agencies. Procurement contracts awarded by a public security agency are most likely to provide access to massive, linkable, personal data, collected for monitoring purposes, while contracts with other agencies likely provide access to less data.[4] Thus, our first proxy for a data-rich contract is one that came from a public security agency, whereas a data-scarce contract is one that did not. We next distinguish among contracts within the set of public security contracts, identifying those that are likely to be especially rich in data. These are contracts with public security agencies possess-

[4]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.

ing greater surveillance capacity, which we measure using prefectural government contracts for surveillance cameras. Thus, our second proxy for a data-rich contract is 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. We prefer this proxy as it allows us to make comparisons *within* a set of very similar public security contracts.

Using these newly constructed datasets, we use a triple differences design to estimate the effect of access to greater amounts of government data on facial recognition AI firms' subsequent software development. Specifically, we compare firms' software releases before and after they receive their first government contract, 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 exploit variation in the type of contract: data-rich or data-scarce. 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. Our evidence is thus consistent with the presence of economies of scope, reflecting crowding-*in* rather than crowding-out. Using our preferred proxy for data-richness, we find that 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.

We provide a range of corroborating evidence for our proposed mechanism of access to government data contributing to product innovation. First, we observe lower bids (even controlling for firm fixed effects) for data-rich contracts, as well as more bidders overall. Second, 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. Finally, we find that firms that produce video facial recognition AI software for the government — which requires access to particularly large amounts of data — exhibit differentially large increases in data-complementary software production, and greater commercial and government AI software production too.

We conclude our empirical analysis by evaluating a range of threats to identification and alternative mechanisms. First, we show that systematic firm selection into receiving contracts at a particular time is unlikely to drive our main results: 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 in their terms and tasks required, potential for learning-by-doing, access to capital, signaling value, associated commercial opportunities, or connections to local government.

Significant microeconomic consequences of economies of scope arising from government data do not necessarily imply that provision of government data would promote *aggregate* innovation or increase welfare. To examine the macroeconomic implications of government data access, we develop 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 procuring a government contract.

We show conditions under which there is a unique balanced growth path (BGP) equilibrium with free-entry of innovators and three types of firms being present:

those producing both government and commercial software using government and private data, those producing private software using private data alone, and those producing non-software. Then, we study how government data access affects innovation and welfare. When commercial software and non-software are sufficiently substitutable, an increase in government data provided to firms increases the BGP rate of innovation and biases private innovation towards data-intensive software. However, the welfare effect is more ambiguous: while government data provision does lead to a direct positive effect on welfare through higher consumption growth, this is offset by a decrease in the level of consumption due to crowding-out by resources used in innovation. Thus, we consider a second-best problem where the state can only choose the level of government data provided to firms in order to maximize household welfare. We find that, even if neither the state nor the representative household derives utility from surveillance, it may be optimal for the state to produce it in order to provide firms with the government data that is generated as a by-product. This is because, in doing so, it can increase the rate of private software innovation and thus consumption growth when there are economies of scope. Importantly, such a policy is only justified when economies of scope are strong enough and, as a result, the increase in the growth rate is sufficiently large to compensate for the crowding-out of resources.

Finally, in three applications, we illustrate the varied ways that data-intensive innovation may be shaped by the state, both directly and indirectly, because of the features of data that we highlight. First, we show that industrial policy in the form of government data provision can be justified on grounds which differ from those that motivate traditional industrial policy. Second, we show that surveillance states' desire to monitor and control their citizens aligns with promoting data-intensive innovation and growth, but may reduce citizen welfare when states' objectives and citizens' preferences do not coincide. Third, we show that regulation limiting government data collection reduces data-intensive innovation and growth but may benefit citizens overall when they value privacy.

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 direct and indirect ways in which data-intensive innovation may be shaped by the state, and identifying two crucial characteristics of data that shape the impact of the state on innovation. Our analysis complements a recent literature studying the effects of specific characteristics of information and data on innovation: Williams (2013) studies the non-excludability of government research on genes (in contrast with the excludability of private sector research); and, Aghion et al. (2017) and Jones and Tonetti (2018) study non-rivalry of data across firms. We instead study economies of scope arising from the sharability of government data across government and commercial applications within a firm.

Our examination of the link between the state and the private sector AI industry builds on 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. Recent research on innovation policy also suggests an important role for the state in encouraging R&D — see Bloom et al. (2019).[5]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. 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

[5]Among others, Howell (2017) shows that the US Department of Energy's funding helps firms to innovate; Azoulay et al. (2018) show that public grants increase patenting by pharmaceutical and biotechnology firms; Moretti et al. (2019) show that defense R&D expenditures of OECD countries crowd in private R&D; and Moser (2005) studies how intellectual property rights shape innovation.

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 (Acemoglu, 1998; Acemoglu et al., 2012; Lewis, 2013; 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 autocratic states (Guriev and Treisman, 2019). In addition, a fundamental aim of AI technology — to make accurate predictions — is aligned with their surveillance and social control agenda (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.[6] 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 em-

[6]The risk of *ex post* taxation or expropriation of entrepreneurs will mean *ex ante* less investment (North et al., 2009; Acemoglu and Robinson, 2012). Threats to elites arising from successful entrepreneurs will mean that elites may *ex ante* tax entrepreneurs to preserve their political rents (Acemoglu and Robinson, 2006). Corruption and other public sectors distortions will also discourage innovation and investment (Shleifer and Vishny, 2002).

phasize, 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.[7]

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.[8] In order to obtain access to these government data, the firm must obtain a contract from the state to produce government software. Government 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 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

[7] A large literature studies the Chinese economy and its spectacular growth in the recent decades (e.g., Song et al., 2011), 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 institutional features that allow China to grow despite the lack of institutional constraints on the Chinese Communist Party. In contrast, our work (along with others, like Bai et al., 2019) identifies a mechanism through which autocratic power can actually promote economic growth.

[8] 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.

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.[9]

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.[10] 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 commercial 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 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

[9] An alternative plausible specification of the technologies is one where government data is not shared across uses per se, but is instead used to train a "base algorithm," which is used as an input to both government and private software. For the purposes of our paper, these two are equivalent.

[10] 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 $C(S_g, S_c, d_g, \omega) < C_g(S_g, 0, d_g, \omega) + C_c(0, S_c, 0, \omega)$.

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 a prototypical setting in which 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 image or video datasets. Second, because the Chinese state collects huge amounts of personal 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."[11] 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."[12]

Consider an example in which a private firm receives a procurement contract to provide facial recognition software 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 could include video from street surveillance cameras as well as 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 being used to train a separate algorithm that results in a commercial AI product, for example, AI software designed for retail firms who may wish to track or detect individual shoppers throughout their stores, and then predict their consumption choices.

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

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

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 the orientation of software toward 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.[13]

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.[14] We extract firms that are categorized as facial recognition AI producers by the database, and we validate the categorization by manually coding firms based on their descriptions and product lists. We complement the *Tianyancha* database with information from *Pitchbook*, a database owned by Morningstar on firms and private capital markets around the world.[15] 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.[16] We also collect an array of firm

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

[14] See Supplementary Figure S.1 for an example entry.

[15] See Supplementary Figure S.2 for an example entry.

[16] 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).

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.[17] 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 the number of bidders for a subset of the contracts. To identify contracts procuring facial recognition AI, we match the contracts with the list of facial recognition AI firms, identifying 26,200 procurement contracts involving at least one facial recognition AI firm. Many firms receive multiple contracts; overall, 1,095 of the 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.[18]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.

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

[17]See Supplementary Figure S.3 for an example contract.

[18]Source: `https://go.aws/37GbAZG`.

use, we aim to be conservative in our measure of commercial software products.

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

To implement the two dimensions of categorization using the RNN model, we manually label 13,000 software products to produce a training corpus. We then use word-embedding to 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.1 shows the summary statistics of the categorization output by customers and by function; and, Appendix Figure A.2 presents the confusion matrix (Type-I and Type-II errors) of the predictions relative to categorization done by humans.[19]

Measuring the quantity of government data to which firms have access We construct two proxies for access to greater amounts of government data. We begin by distinguishing among government contract awarding agencies. Procurement contracts awarded by a public security agency are most likely to provide access to massive, linkable, personal data, collected for monitoring purposes, while contracts with other agencies likely provide access to less data. Take, as an example from our dataset, a public security contract signed between an AI firm and a municipal police department in Heilongjiang Province to "increase the capacity of its identity information collection system" on August 29th, 2018. The contract spec-

[19]Supplementary Table S.1 presents the top words (in terms of frequency) used for the categorization. Supplementary Figure S.4 presents the density plots of the algorithm's category predictions.

ifies that the AI firm shall provide a facial recognition system that can store and analyze at least 30 million facial images — a substantial amount of data to which the firm obtains access. In contrast, consider a non-public security contract in our dataset signed between an AI firm and a provincial bank in Gansu Province to "establish its facial recognition system" on November 20th, 2018. The system is aimed at providing identification services for the bank's clients, suggesting that the AI firm obtains access to a relatively small amount of data (i.e., identified faces) compared to a public security contract.

Our first empirical definition of a data-rich contract is a contract with a public security agency, the effects of which we compare to those of data-scarce procurement contracts, awarded by government agencies unrelated to public security (e.g., contracts with schools to monitor cheating during exams). Such non-public security contracts indicate a firm's relationship with the government, but do not provide access to large amounts of personally identified facial data.[20]

Our measure of public security contracts is comprehensive. We capture the following four types of 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.

We identify 28,023 public security procurement contracts involving at least one facial recognition AI firm. Many firms receive multiple contracts: 7.2% (12.6%) of the facial recognition AI firms in our dataset receive at least one public security (non-public security) procurement contract; and 5.2% of the facial recognition AI firms receive at least one contract of each type.

We next distinguish among contracts *within* the set of public security contracts, identifying those that are likely to be especially rich in data for facial recognition AI firms. In particular, we identify contracts with public security agencies possessing greater video surveillance capacity, which we measure using 5,837 prefectural

[20]We identify 410,510 public security contracts in total. Both the public security and non-public security contracts have steadily increased since 2013. See Supplementary Figure S.5.

Figure 1: Circle size indicates the number of first AI contracts awarded in the prefecture. Circle shading indicates the fraction of first AI contracts that were data-rich or data-scarce, where the within-prefecture variation comes from changes in the number of surveillance cameras over time.

government contracts for surveillance cameras.[21] We sum the number of cameras procured in each prefecture up to a certain date and divide this by the prefecture's population to form a time-varying measure of the video surveillance capacity of a particular prefecture.[22] Our second — and preferred — 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. Therefore, this captures the amount of *identifiable* data that a firm may gain access to.[23] Figure 1 shows the distribution of data-rich and data-scarce contracts across prefectures according to this second, preferred definition.[24] We compare the effects of these data-rich contracts to data-scarce public security contracts, now defined as contracts awarded by a public security agency, but located in a prefecture that has below-median surveillance capacity at the time the contract was awarded. We prefer this definition of a data-rich contract given the fineness of the comparison within a set of firms that selected into a similar set of public security contracts.

Summary statistics Appendix Table A.2 presents summary statistics describing

[21]There are on average 77 contracts per prefecture. In Supplementary Figure S.6, we present a time series plot of the number of cameras in our data over time.

[22]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; source: `https://bit.ly/3dqdjU0`.

[23]Note that the existence of a national ID system in China likely implies that there is 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.

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

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 data-rich and data-scarce contracts will be crucial to identify the effects of the contracts. Interestingly, some patterns of selection into contracts that are data-rich differ depending on the definition used: for example, firms receiving public security contracts are better capitalized than firms receiving non-public security contracts (40 vs. 13 million USD), 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). This suggests that simple selection stories will not easily account for effects of data-rich contracts seen along both margins of comparison.

Appendix Table A.3 presents summary statistics describing the contracts procuring AI services in our sample.[25] Data-scarce and data-rich contracts differ on dimensions other than in the quantity of data to which firms receive access, so accounting for alternative mechanisms (other than data provision) through which data-rich contracts might affect software production will be crucial to identifying the causal effects of interest. However, it is worth noting that the differences observed between data-rich and data-scarce contracts often reverse depending on which definition of data-rich is used. For example, public security contracts are on average issued by a lower administrative unit than non-public security contracts (28% vs. 34% by provincial level or above), but public security contracts issued in prefectures with above-median surveillance capacity are issued by a higher administrative unit than public security contracts issued in prefectures with below-median surveillance capacity (31% vs. 14% by provincial level or above). Finding consistent effects of data-rich contracts across definitions will argue against simple alternative hypotheses regarding unobserved contract characteristics.

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 and inno-

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

vation. The empirical strategy exploits variation across time and across firms in the receipt of a government contract, and across types of government contracts that firms receive. Specifically, as in an event study design, we compare firms' outcomes — their software releases — before and after they receive their first government 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 exploit variation in the type of the government contract received.

We test whether firms receiving data-rich 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 contract, and $Data_i$, a dummy variable indicating whether the firm's first contract was data rich, as defined above.[26]

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.[27] 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.[28]

[26] We focus on the effect of the initial contract, because the receipt of subsequent contracts is endogenous to firms' performance in their initial contracts.

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

[28] We cluster standard errors at the mother firm-level to be conservative; clustering standard

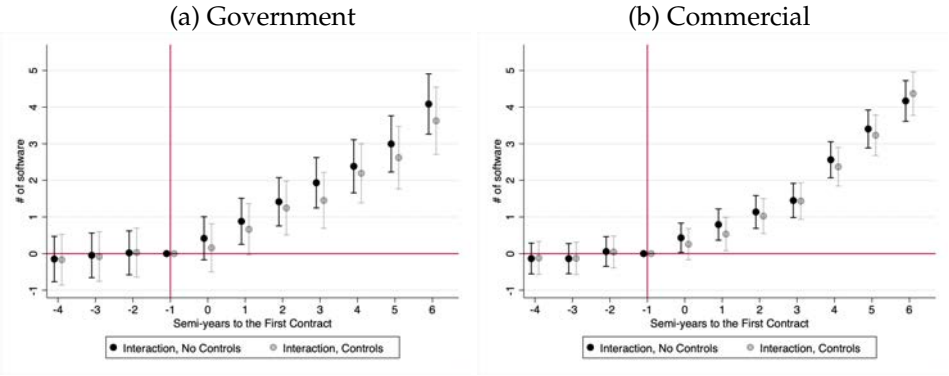
Our empirical strategy allows us to address important threats to identification. A particular concern is non-random assignment of contracts to firms. We account for fixed firm characteristics that may determine selection into data-rich contracts as well software production by including a full set of 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

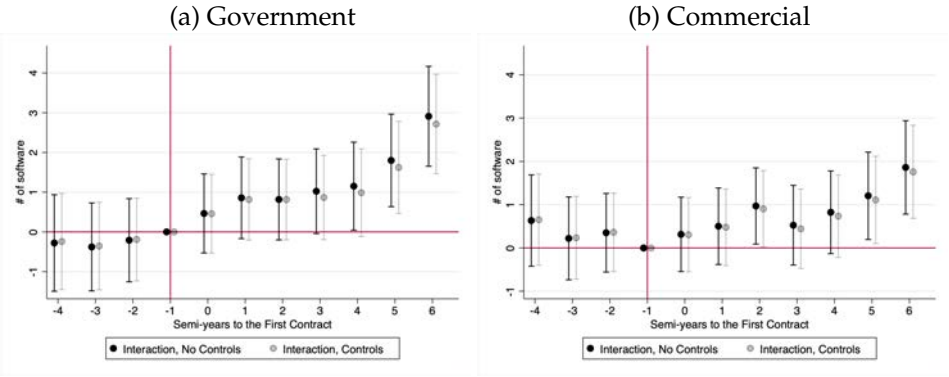
5.2.1 Baseline estimates and robustness checks

We first estimate our baseline specification, comparing the effects of public security contracts to non-public security contracts on firms' production of software. In Figure 2, Panel A, we plot the coefficients β_{1T}, describing the *differential* software production around the time when a public security contract was received, relative to a non-public security contract (all coefficients are presented in Appendix Table A.5). We show 95% confidence intervals for all coefficients, from models with and without controls ($\sum_T T_{it} X_i$). In Panel A(a), one can see that receipt of a public security contract is associated with differentially more government software production than receipt of a non-public security contract. Suggesting a causal interpretation of the effect of a public security contract, 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. In Panel A(b), one can see that receipt of a public security contract is also

errors at the firm level allows us to make even more precise inferences.



Panel A: Public security vs. non-public security contracts



Panel B: Public security contracts with high vs. low surveillance capacity prefectures

Figure 2: Differential software development intended for government (left column) or for commercial uses (right column), resulting from data-rich contracts, relative to data-scarce contracts, controlling for firm and time period fixed effects. Panel A defines data-rich contracts as all public security contracts. Panel B defines them 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.

associated with differentially more *commercial* software production than receipt of a non-public security contract. Again supporting a causal interpretation of the effect of a public security contract, 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.

We next, in Figure 2, Panel B, plot regression coefficients analogous to those in Panel A, but now considering variation in data-richness *within* the set of public security contracts. Specifically, we compare 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). All coefficients are presented in Appendix Table A.6. This is our preferred proxy of

data-richness as it accounts for two potential concerns about our previous comparison between public and non-public security contracts. First, that firm selection into public and non-public security contracts may be different. Second, that public and non-public security contracts may differ beyond the quantity of government data the firms can access (e.g., the type of government software developed and its production process could also differ). We focus on our preferred proxy for data-richness in our subsequent empirical analyses, but results are qualitatively identical comparing public security and non-public security contracts instead.

In Figure 2, Panel B(a), we examine government software production. 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. Suggesting a causal interpretation of the effect of a data-rich public security contract, 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 does not affect our findings. In Figure 2, Panel B(b), one sees that receipt of a data-rich public security contract is also associated with differentially more *commercial* software production than receipt of a data-scarce public security contract. Again supporting 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. In terms of magnitudes, we see in Figure 2, Panel B, that 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.

Interpretation As discussed in Section 3, the results presented above indicate economies of scope in AI innovation arising from government 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, 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 all types of govern-

ment contracts on commercial software production, and differentially larger effects of data-rich contracts than data-scarce ones (see Appendix Figures A.3– A.4).

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 (either non-public security contracts or public security contracts in prefectures with below-median surveillance capacity), we account for time-varying selection into receiving either a government or 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 government 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 5.2.2 and 5.2.3, we provide further evidence of the importance of government data.

Robustness Given the complex process of constructing our dataset, it is important to note that our findings are robust to varying several salient dimensions of our analysis (see Appendix Table A.7). First, our results are robust to adjustments of the key RNN LSTM classification algorithm parameter choices — timestep, embedding, and node (see Panel A). Second, the results are robust to adjustments of the LSTM classification threshold (see Panel B). Third, our results are robust to considering a balanced panel of firms within a narrow window, and to expanding the window of time around the receipt of the first contract that we study (see Panel C). Finally, our results are robust to adjusting our classification of (data-rich) public security contracts to exclude any ambiguous government agencies (e.g., contracts with the government headquarters could be meant to provide security services just for the government office building; see Panel D).

5.2.2 Additional evidence of the importance of data as an input

Our proposed mechanism of economies of scope arising from government data suggests that 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.5): 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 $= 0.05$).

Under our proposed mechanism, 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 Appendix Figure A.6, Panel A, we present estimates from the same specification as in Panel B of Figure 2, but now considering the outcome of data-complementary software products. One can see that the data-complementary software production *differentially* increases after the receipt of a data-rich public security contract.[29] 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.[30]

Our proposed mechanism suggests that access to government data will not only increase the quantity of software production, but also the quality. While we cannot directly observe the quality of software produced, we can test whether

[29] 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 (see Appendix Figure A.6, Panel A, left and middle columns).

[30] The production of data-complementary software can be seen as an alternative empirical proxy for firms' receiving access to particularly large quantities of data. Analogous to our previous comparison between high and low surveillance capacity public security contracts, one would expect differentially more government and commercial AI software production among firms that produced data-complementary software after receiving a public security contract. In Appendix Figure A.7, one can see that public security contracts that led to data-complementary software production within the first year of the contracts were associated with differentially more government *and* commercial software production.

data-rich contracts lead to increased production of the most demanding form of facial recognition AI: that using video. Indeed, we find significantly greater video facial recognition AI software production following receipt of a data-rich contract (see Appendix Figure A.8).

A final set of tests arises from an examination of 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.[31] 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.9 that indeed they do. Moreover, we note that the magnitudes of the coefficients when considering the post-contract production of government video AI as a proxy for the data-richness of the contract are nearly double those using our other proxies, consistent with the idea that video AI software is particularly data-intensive.

A range of tests, exploiting multiple margins of variation in access to government data, all point in the same direction: beyond other mechanisms through which contracts may affect output, access to government data plays a crucial role.

5.2.3 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, including alternative sources of economies of scope. 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, defining data-rich contracts as those public security contracts from prefectures with above median surveillance capacity. The figure plots these estimates specification-by-specification. We also present more complete sets of estimates in Appendix Tables A.5 to A.8, including those resulting from defining data-rich contracts as public security contracts instead.

[31] 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 Figure A.9 and Figure A.6, Panel B.

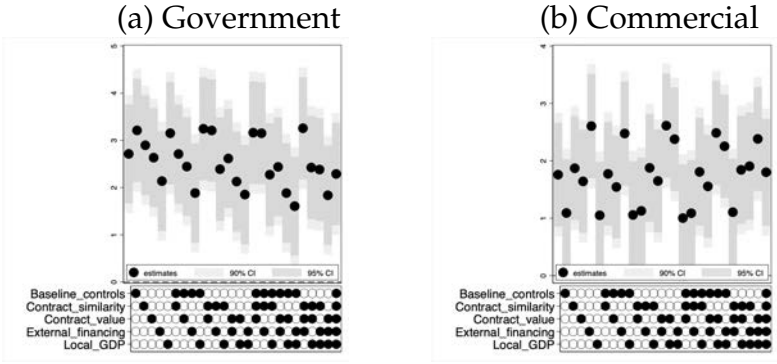


Figure 3: Software development intended for government (left, (a)) and commercial (right, (b)) use relative to the time of receiving initial procurement contract. Figure shows coefficient on the interaction for high surveillance capacity 3 years after public security contract receipt, controlling for firm and time period fixed effects and adding various controls. Solid dots indicate significance at the 10% level or better. Results presented in Appendix Table A.6.

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 (high or low capacity), we calculate the vector distance between the language of each public security contract in our dataset 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.8, 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. In particular, we evaluate two such alternative hypothesis. First, 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.[32] Two 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 (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.8, Panel B).

Second, systematic differences in the type of government software may be more of a concern when we compare public and non-public security contracts, since the production processes for the associated government software may be different. Yet, for our preferred comparison within the set of public security contracts, we view such positive correlation as much less likely. In fact, 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, suggesting that the underlying production processes and types of government software should be similar as well.

[32]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.

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.8, 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.8, 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 receiving a government contract receive additional benefits from local industrial policy, or attract additional external funding, human capital, or customers, all of which contribute to the production of software. To test whether the signaling value of data-rich contracts accounts for our findings, we first 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.8, Panel C, one can see that within this sample of first contracts 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.8, 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.8, 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.8, 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 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. We next explore the macroeconomic implications of these findings.

6 Macro implications of firms' government data access

In our empirical analysis of Section 5, we have observed some of the firm-level consequences of access to government data: an increase in government data available to firms increases their data-intensive innovation. However, this evidence does not imply that such policy will shift the *aggregate* direction of innovation or the economy's growth rate. There are two main reasons why the microeconomic and macroeconomic implications may diverge. The first is that increases in innovation by firms accessing government data may crowd-out resources from other innovating firms. The second is that, in general equilibrium, relative prices may change, thus affecting innovation as well. Moreover, even if the economy's growth rate

did increase, it would not necessarily imply that increasing firms' access to government data would increase welfare. A higher consumption *growth rate* is offset by a lower *level* of consumption due to crowding-out of resources by innovation and government data production.

Thus, in this section, we examine how access to government data affects the direction of innovation, growth, and welfare in data-intensive economies, with these considerations taken into account. To do so, we build a directed technical change model (Acemoglu, 2002) with data as an input and economies of scope.

6.1 A directed technical change model with data as an input

Model overview We model an economy in which firms innovate to develop and supply 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.[33] 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 gov-

[33]This corresponds, for instance, to information collected from consumers when performing online transactions. Note that, with this setup, we are ignoring interesting issues regarding how to allocate private data property rights between firms and consumers.

ernment 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 profits 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:

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

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.[34] 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

[34]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.

and d_p, as well as intermediates x_c to produce a variety of quality q_c.

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.[35] 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 normalized to 1) and that varieties never depreciate.[36]

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.[37] Then, given total R&D spending R_i for sector i, new varieties accumulate according to:

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

[35]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.

[36]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.

[37]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. A more realistic model would incorporate researchers too, as in a "knowledge-based R&D model" (Acemoglu, 1998).

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 both types of software firms present requires:[38]

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

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

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:

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

[38] 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.

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:

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

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 surveillance: specifically, one unit of surveillance, G, produces κ_g units of government data.[39] 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.

[39] In our empirical context, this government data could correspond, for example, to the video feed from street cameras or individual administrative records. These are themselves produced as a consequence of the surveillance and other activities of governments.

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. Suppose it can mine $\kappa_p Y$ units of data out of Y, then the supply of private data is:[40]

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

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

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.[41] Proposition 1

[40] 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.

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

in Appendix A.1 lays out sufficient conditions for a BGP to exist and be unique where all types of firms are present.

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 $\omega = \frac{q_c(0,p_c,d_p)^{1-\frac{1}{\chi}}}{q_c(0,p_c,d_p)^{1-\frac{1}{\chi}} + q_c(\bar{d}_g,p_c,d_p)^{1-\frac{1}{\chi}}}$.

6.2 The consequences of government data provision

We now focus on two questions, one positive and one normative: first, how does government data provision affect the rate and direction of innovation? Second, how does government data provision affect welfare?

How does government data provision affect the rate and direction of innovation? The next theorem shows the conditions under which policies that directly provide more government data to innovating firms increase the economy's growth rate and bias the direction of private innovation towards data-intensive software.

Theorem 1 (Government data provision and innovation) *Assume the sufficient conditions in Proposition 1 for a unique BGP equilibrium to exist hold. Then, an increase in government data provided to firms ($\bar{d}_g$) will increase the rate of innovation (η). Moreover, if $\epsilon \geq \frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}$, 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 profits earned by firms using government data will drive up the return on investment (r) under free-entry and,

therefore, induce higher R&D spending and increase 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.

Finally, we establish an equivalence between state policy choices which anticipates some of the results in Section 7: the indirect effects of a state's choice of surveillance levels and data collection will, in a BGP, be analogous to the direct effects of government data provision. See Supplementary Material B.1 for a proof.

Remark 1 *In a BGP equilibrium, the consequences of an increase in $\bar{d}_g$ are equivalent to those arising from an increase in surveillance spending $p_g G/Y$ or aggregate government data D_g/Y as a share of final good output.*

How does government data provision affect welfare? 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 increase η *per se*. The appropriate objective for a benevolent state is to maximize household utility. Assuming $\rho - \eta(1-\theta) > 0$ on a BGP (i.e., utility is bounded), the present discounted utility of the representative household is (aside from the initial level of output which we normalize to 1):

$$U = \frac{1}{1-\theta}\left(\frac{C}{Y}\right)^{1-\theta}\frac{1}{\rho-\eta(1-\theta)}.$$

The increase in η leads to a direct positive effect on welfare, since the growth rate of consumption is higher. But, from the aggregate resource constraint (shown below), we see that there are two forces that may *offset* such increase by decreasing the consumption to output ratio $\frac{C}{Y}$ (and therefore welfare) following an increase in $\bar{d}_g$:

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

First, the crowding-out of resources from consumption that are used instead for creating new varieties (i.e., innovation). Second, the crowding-out of resources

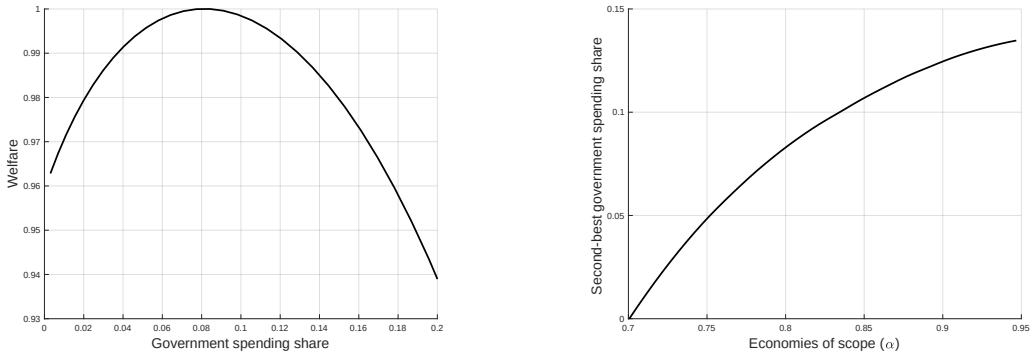


Figure 4: Left panel: government data provision and welfare; Right panel: Economies of scope and second-best government data provision.

from consumption that are used instead as intermediates inputs in surveillance.

Given this discussion, we next consider a second-best problem where the government chooses the level of government data provision to maximize household welfare.[42] Figure 4 shows how welfare changes when the share of government spending in total output $\frac{p_g G}{Y+p_g G}$ changes. These changes are brought about by different levels of government data provision $\bar{d}_g$. We report welfare in consumption equivalent units relative to the maximum attainable welfare. To make the analysis transparent, the benchmark parameterization underlying the figure is such that: *(i)* the economy is symmetric in the sense that the direction of innovation is unbiased ($\frac{\tilde{N}_c}{N_z} = \frac{N_g}{N_z} = 1$), 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 are identical ($\bar{d}_g = d_p(\bar{d}_g, p_c, p_d)$); and, *(ii)* economies of scope (as governed by α) are consistent with our benchmark estimates from the empirical section (to be precise, the relative elasticity of commercial to government software production of around two-thirds implies $\alpha = 0.82$).[43] Then, we vary the level of government data provision $\bar{d}_g$ from this benchmark parameterization, keeping all other parameters fixed.

We view this numerical exercise (and the ones in the following sections) as qualitatively illustrating the forces at play. While we have shown direct empirical evidence that disciplines α, a comprehensive quantitative assessment would further require measuring a number of other parameters about which we still have a large degree of uncertainty (e.g., the substitutability of data with other inputs or the substitutability between commercial software and non-software).

In Figure 4, left panel, one sees that, given our parameterization of the model,

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

[43] See Supplementary Material B.3 for a more detailed description of the calibration.

the second-best government data provision results in a government spending share of 8%. Moreover, one can see that deviations from this second-best can be rather costly. For example, when government data is relatively scarce and the government spending share is only 2%, then welfare is about 2% lower in consumption equivalent units. The reason is that the growth rate is lower, and this is not sufficiently compensated by less crowding-out of resources from consumption by innovation. The opposite is true when government data provision is too generous.

These results beg the question as to what determines the welfare maximizing government data provision. Is it always the case that an interior solution exists? Or, would it sometimes be optimal for the state not to provide government data at all? To answer this, Figure 4, right panel, shows how the welfare maximizing government spending share changes as economies of scope become stronger. We find that when α is below 0.7 then it is never optimal for the state to supply 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 or surveillance. As economies of scope become greater, so does the second-best government spending share, because a higher level of government data provision to firms causes larger changes in the economy's growth rate.

7 Roles of the state in data-intensive economies

In this section, we present three applications which illustrate the varied ways that data-intensive innovation may be shaped by the state, both directly and indirectly.

States' choice of industrial policy Traditional forms of industrial policy entail giving direct production subsidies to a sector or subsidizing a key input. The stated goal of such policies is often to shift the relative size of sectors (and/or the direction of innovation) to correct for market failures. Alternatively, states sometimes directly provide key inputs that are used by private firms. These include, for example, infrastructure — such as transportation, water, or electric power — as well as public services that increase worker productivity — such as education or health.[44]

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

Our evidence and model suggest another justification for industrial policy in the age of data-intensive innovation. Because states are key collectors of data and government data gives rise to economies of scope, in Section 6.2 we have shown that it may be optimal to directly provide such data to data-intensive software producers when they contract with the government.[45] The justification is even stronger when government services that produce data as a by-product (like surveillance) are also directly valued by either the state or households.

Our model also suggests that differences in production technologies across data-intensive economies can have important effects on industrial policies. For instance, Figure 4 has shown the consequences of variation in economies of scope, which themselves may be due to differences in sectoral composition across economies.

Finally, the next two applications show that, even without intending to do so, different state policies may also have important industrial policy components, echoing the arguments of Rodrik (2007).

States' choice of surveillance level All states engage in citizen monitoring for the preservation of public security, potentially generating massive surveillance datasets. At the extreme are autocratic states that aim to monitor and control their populations to maintain power (Guriev and Treisman, 2019). In the modern world, this need to monitor is likely to produce substantially greater data collection and data analysis — particularly using AI. Indeed, AI has been described by the *Wall Street Journal* as part of the "autocrat's new tool kit."[46] China is one 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.[47]

Our model and empirical results 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 the government data (as a by-product) that fuels innovation.

We consider an extension of our model where the flow utility is:

$$u(C) + \delta G. \tag{21}$$

[45] While we do not observe this in our empirical setting, a policy that provides government data to all firms, not just those contracting with the state, would further foster innovation in our model.

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

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

This captures the social welfare function of a state that values both household utility and also G directly. For example, $\delta > 0$ can capture an autocratic state wanting higher G to better monitor the population or, alternatively, a representative household that cares about security provided by the government in democracies facing security threats.[48] Differences in δ across states will result in differences in government spending and, as a result, in growth rates and the bias of private innovation.

More concretely, consider the following thought experiment. Imagine that the only differences between the US and Chinese economies was their δ. Moreover, imagine that this fully explains why government spending on domestic security was 40% lower in the US than China in 2018.[49] We assume that the symmetric economy associated with our benchmark calibration was China. Holding all else fixed, we ask: what are the consequences of decreasing surveillance spending ($\frac{p_g G}{Y+p_g G}$) by 40%? We find that the annual growth rate decreases from the benchmark 6% to 4.7% and the bias of innovation decreases from the benchmark 1 to 0.84.

These results show that surveillance states' preferences for monitoring and controlling their population may result in an inherent advantage in data-intensive innovation by expanding surveillance spending and the provision of government data. While a previous literature (discussed in Section 2) has pointed out that more autocratic regimes may impose a "tax" on private innovation through the hold up or expropriation of entrepreneurs, our findings suggest that this autocratic tax may be offset by surveillance states' advantage in data-intensive innovation. Note, however, that optimal government data provision to firms and surveillance levels chosen by the *state* could be very different from those preferred by *citizens* — states and citizens may have different values of δ. Thus, surveillance states may promote data-intensive innovation and growth, but significantly reduce citizen welfare.

States' choice of privacy protection States not only collect and hold data, but also regulate the collection and exchange thereof. This regulation often reflects citizen norms regarding privacy, as many individuals express discomfort when private data — especially government data — are collected and commoditized.

We focus here on restrictions on the state's collection and sharing of data. Our model suggests that the expression of norms regarding privacy (ultimately re-

[48] Even selfish autocrats may value $u(C)$ if the probability of staying in power increases with it.

[49] US spending was 0.8% of GDP vis a vis 1.32% of GDP for China. Sources: `https://bit.ly/3hdzxeO`, and `https://bit.ly/3aDw32N`.

flected in regulation) can significantly affect the rate and direction of innovation. Consider an extension of our model where the flow utility is:

$$u(C) - \varphi D_g. \tag{22}$$

A positive φ captures households' distate towards government data production and data sharing. If households can enforce these preferences through regulation, then a benevolent state would produce a lower level of aggregate government data in a BGP (i.e., a lower D_g/Y). As a result, if such preferences vary across economies, then the rate and direction of innovation will vary as well. However, reduced government collection of data could be welfare-enhancing for citizens who value privacy, despite lower growth rates and less data-intensive innovation.

More concretely, we engage in the following thought experiment. Imagine that the symmetric economy associated with our benchmark calibration was again China. What would be the consequences of decreasing the amount of government data to the level observed in Germany? Specifically, we consider decreasing D_g/Y by 57% across BGP equilibria, which corresponds to the decrease from the number of cameras in China (14.36 per 100 residents) to Germany (6.27 per 100 residents) in 2018.[50] We find that the annual growth rate (η) decreases from the benchmark 6% to 4% and the bias of innovation (n_c) decreases from the benchmark 1 to 0.76.

8 Conclusion

In this paper, we analyzed direct and indirect ways in which data-intensive innovation may be shaped by the state, highlighting two features of data as an input: (i) historically, states have been key collectors of data, and (ii) data is sharable across multiple uses within firms, giving rise to economies of scope when firms can access government data and produce data-intensive software for government and commercial uses.

Our analysis suggests several directions for future research. First, we have provided a theoretical justification for government data provision as a policy to promote innovation, and evidence on one determinant of its consequences: economies of scope. Yet, many uncertainties remain about the implications of this policy, and a comprehensive quantitative assessment requires further measurement. For

[50] Source: `https://bit.ly/3gdoL6M`.

example, we know little about the substitutability of data with other inputs or the technologies for supplying and collecting data. Moreover, studying a broader range of countries and data-intensive technologies — e.g., for health care or mapping — will help us determine whether government data is as important elsewhere as it is in China's facial recognition AI industry.

Second, we have studied the consequences of *government* data collection and provision to firms. One would also like to study the implications of *private* data collection and dissemination. Private firms' possession of large datasets (e.g., those collected by Alibaba, Baidu, Facebook, or Google) should also generate significant economies of scope. Our analysis suggests that firm access to large private datasets (or lack thereof) will have important consequences for innovation, competition, and growth (see Jones and Tonetti, 2018). Empirically identifying such effects is an important area for future work.

Finally, our analysis sheds new light on interrelationships among innovation, political institutions, and culture that require further study (see, e.g., Benabou et al., 2015; Besley and Persson, 2019). Our work suggests that the alignment between data-intensive innovation and the Chinese state's surveillance interests, as well as permissive privacy norms, can help explain China's rise to pre-eminence in AI. However, we find that the normative implications of greater data-intensive innovation in surveillance states are complex, with higher economic growth potentially coming at a significant welfare cost to citizens. More research is therefore needed to understand the role of the state in determining economic, political, and social outcomes 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.

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.

Azoulay, Pierre, 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.

Benabou, Ronald, Davide Ticchi, and Andrea Vindigni, "Religion and Innovation," *American Economic Review*, May 2015, *105* (5), 346–351.

Besley, Timothy and Torsten Persson, "The Dynamics of Environmental Politics and Values," *Working Paper*, May 2019, pp. 1–38.

Bloom, Nicholas, John Van Reenen, and Heidi L Williams, "A Toolkit of Policies to Promote Innovation," *Journal of Economic Perspectives*, 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.

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.

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.

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.

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

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

Lewis, Ethan, "Immigration and Production Technology," *Annual Review of Economics*, August 2013, *5* (1), 165–191.

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.

North, Douglass C, John Joseph Wallis, and Barry R Weingast, *Violence and Social Orders*, 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.

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

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.

Sejnowski, Terrence J, *The Deep Learning Revolution*, MIT Press, October 2018.

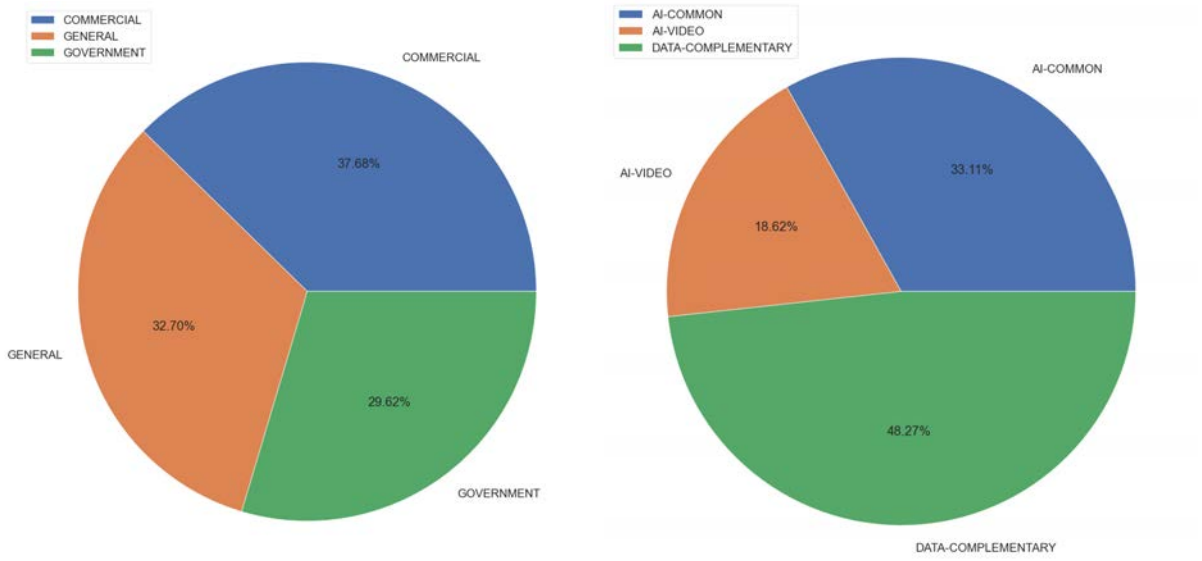
Shleifer, Andrei and Robert W Vishny, *The Grabbing Hand* Government Pathologies and Their Cures, Harvard University Press, 2002.

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

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.

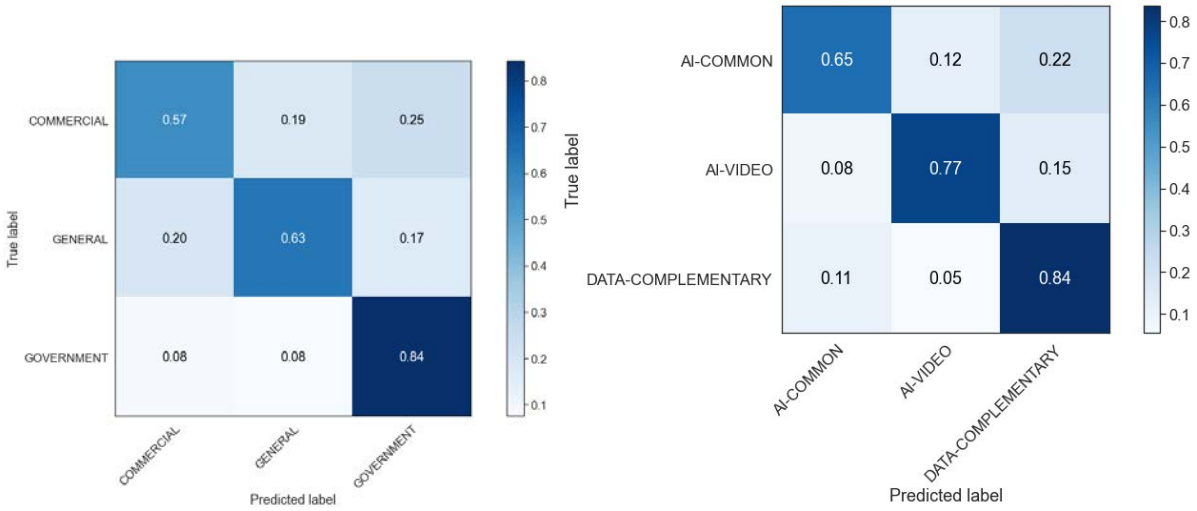
Zuboff, Shoshana, *The Age of Surveillance Capitalism* The Fight for a Human Future at the New Frontier of Power, PublicAffairs, January 2019.



(a) Customers

(b) Function

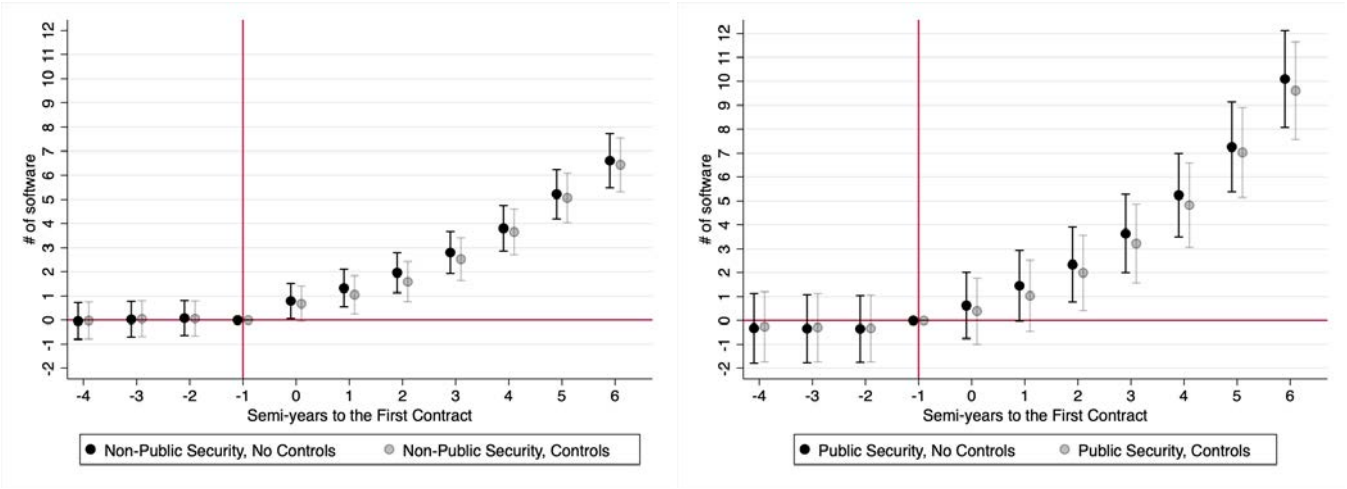
Figure A.1: Summary statistics of categorization outcomes for software categorizations based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization by customers; bottom panel shows categorization by function.



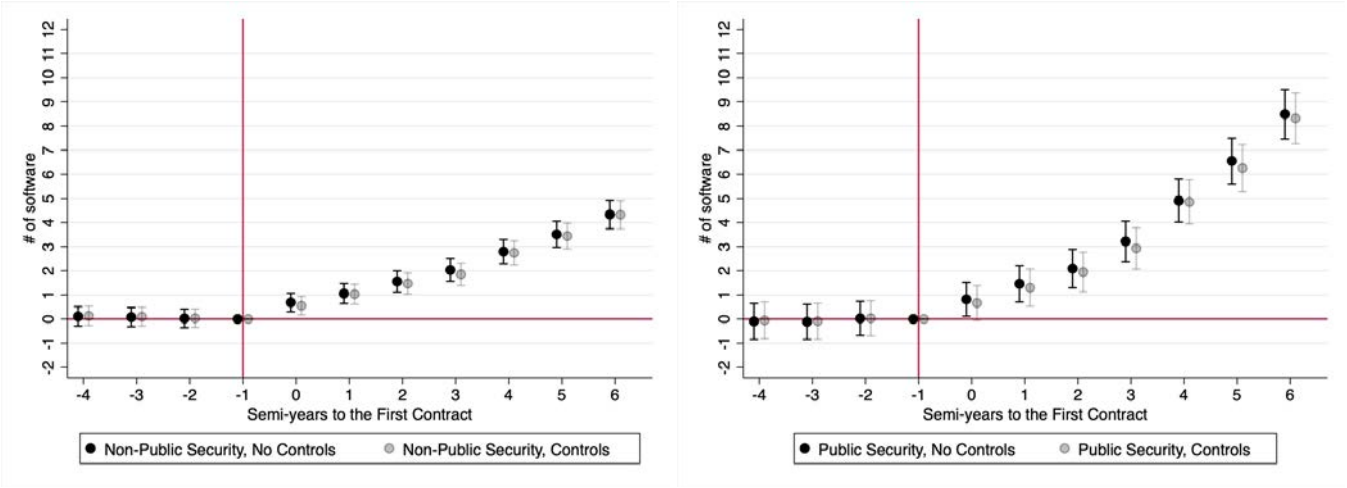
(a) Customers

(b) Functions

Figure A.2: 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.

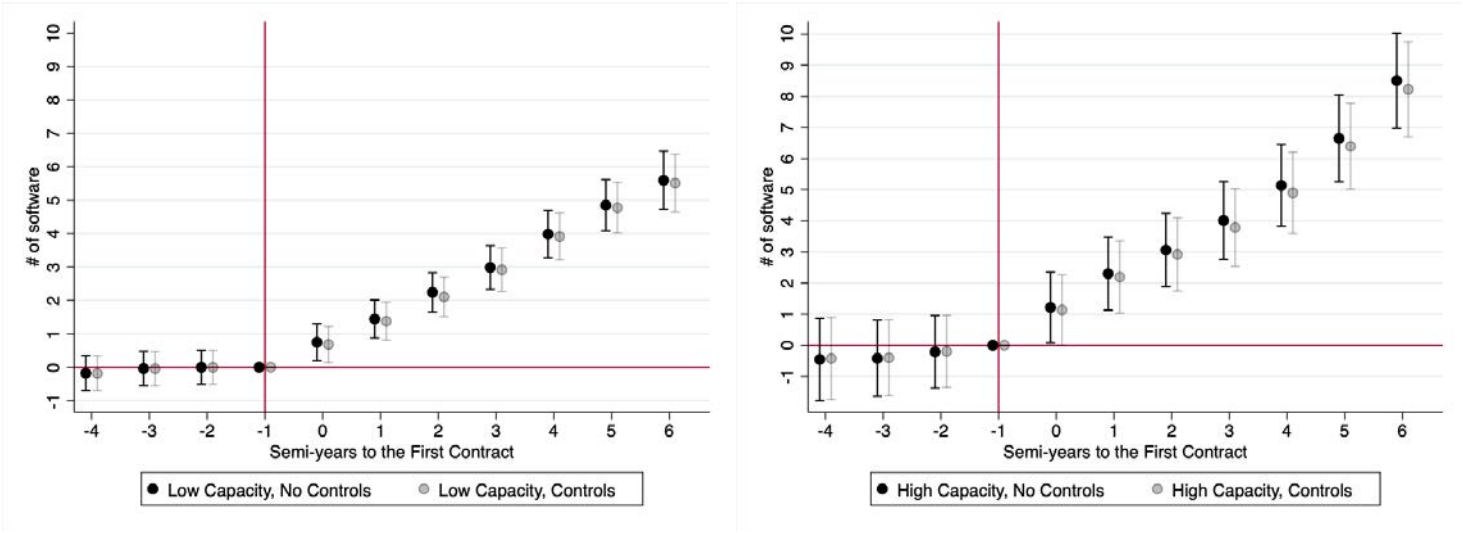


(a) Government

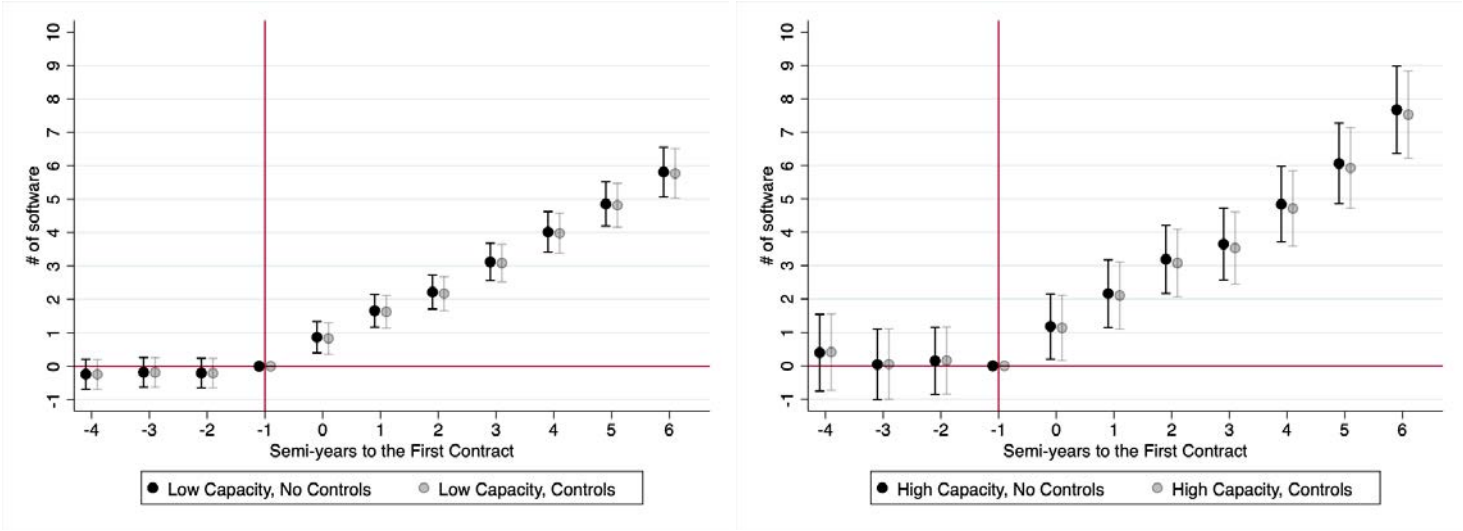


(b) Commercial

Figure A.3: Software development intended for government (Panel A) or for commercial uses (Panel B), resulting from public security contracts (right column) and non-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.



(a) Government



(b) Commercial

Figure A.4: 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.

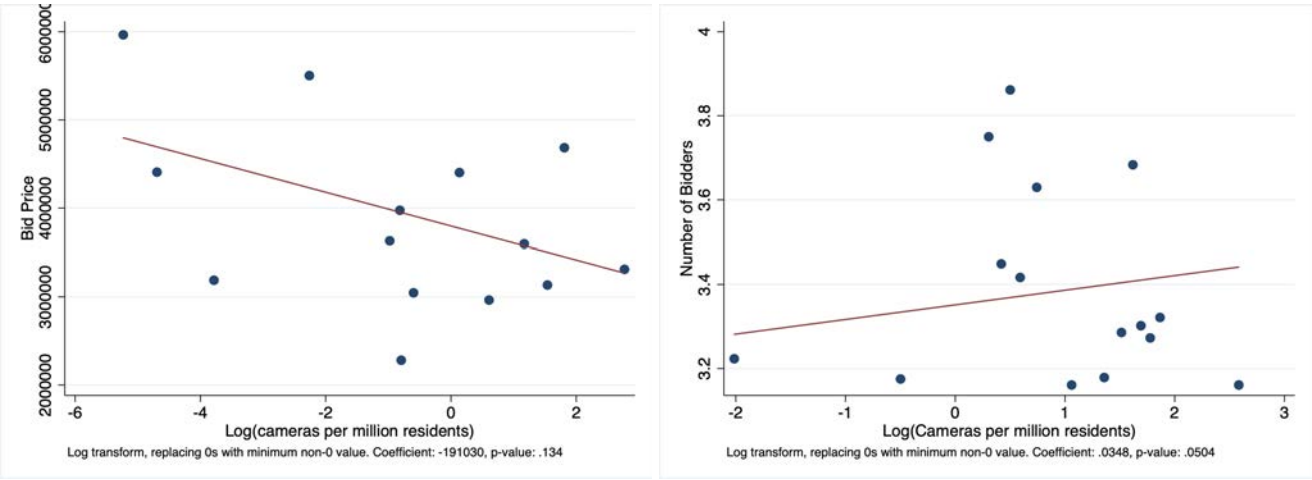
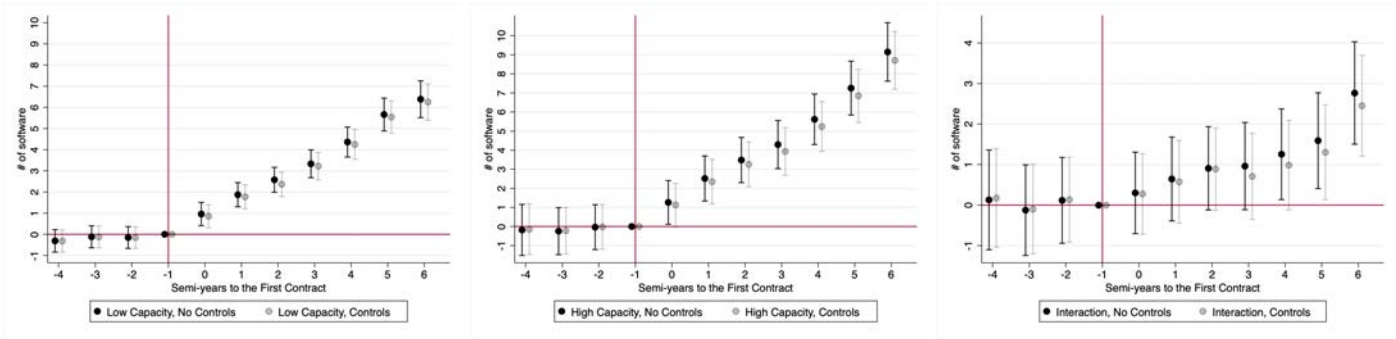
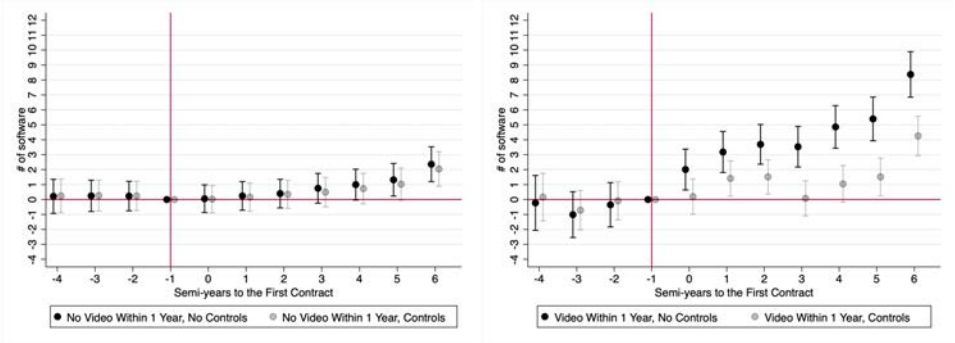


Figure A.5: Binned scatterplots of size of bid versus prefecture surveillance capacity, conditional on company fixed effects (left); and of number of bidders versus prefecture surveillance capacity (right).



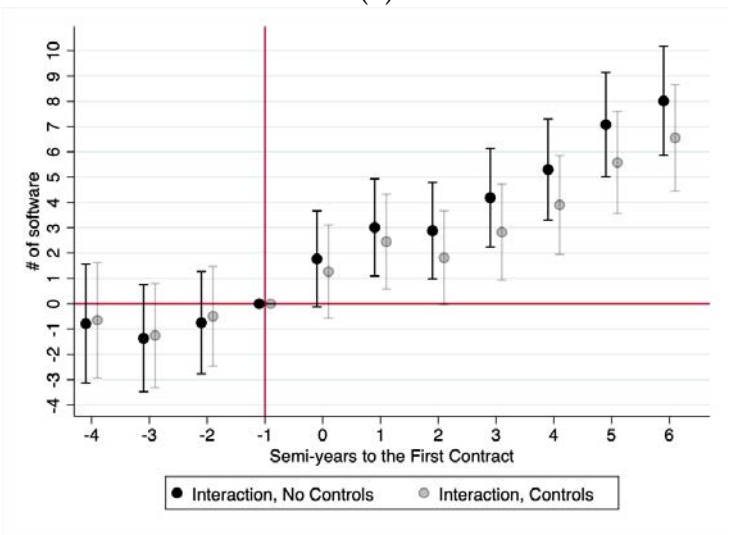
Panel A: Data-complementary, split by surveillance capacity



Panel B: Data-complementary, split by AI video production in year 1

Figure A.6: Panel A: Data-complementary software production resulting from data-scarce contracts (right column), data-rich contracts (middle column), and the difference between data-rich and data-scarce 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: 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) Government



(b) Commercial

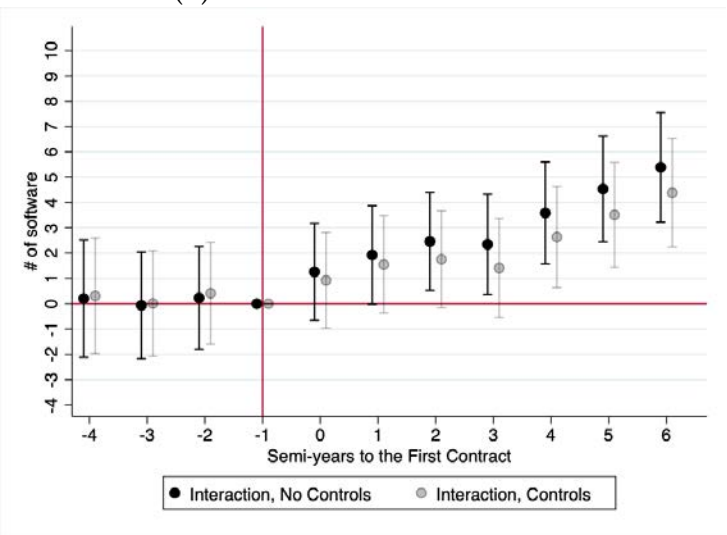
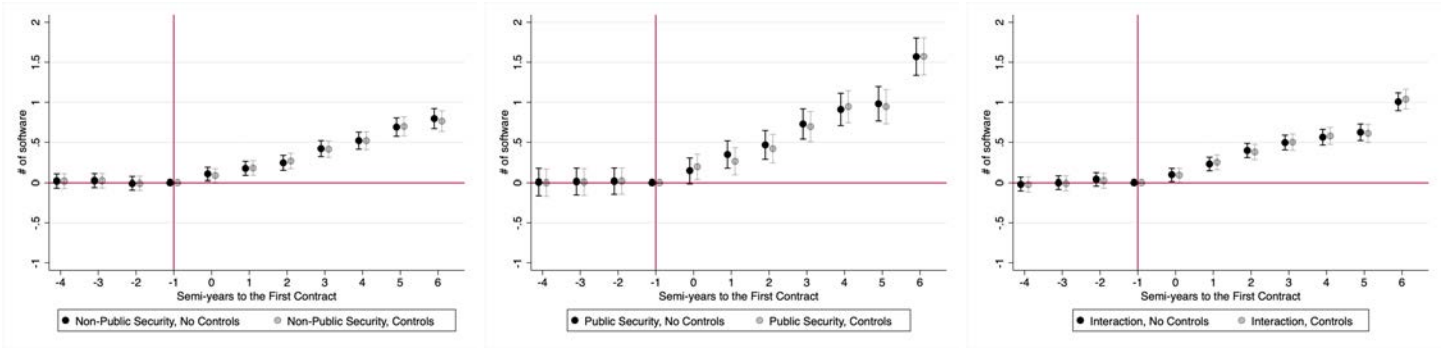
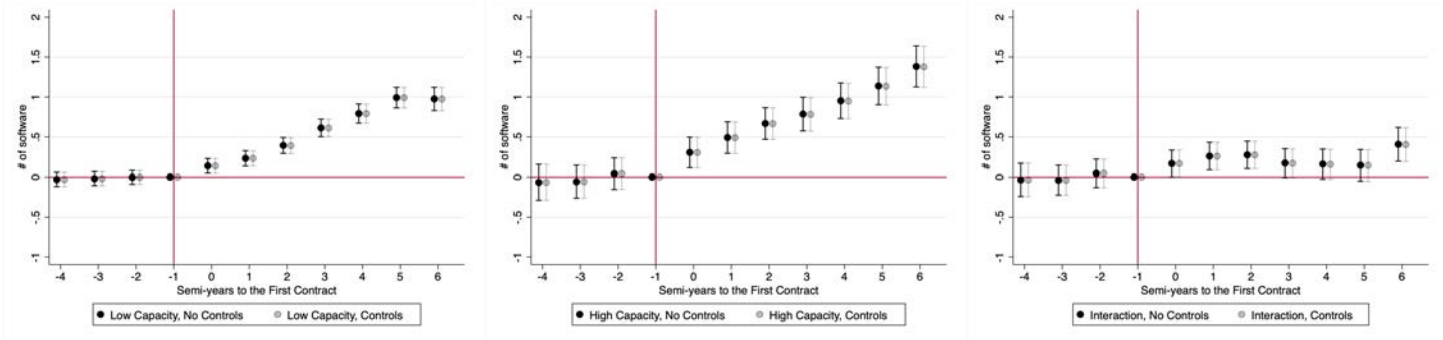


Figure A.7: Differential software development intended for government (left column) or for commercial uses (right column), resulting from public security contracts that led to data-complementary software production within 1 year, relative to public security contracts that did not, controlling for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.



Panel A: Facial recognition AI involving video, public security vs. non-public security contracts



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

Figure A.8: Facial recognition software development that involves video (N-to-N matching). Panel A: Results from non-public security contracts (left column), public security contracts (middle column), and the difference (right column). Panel B: Results 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.

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

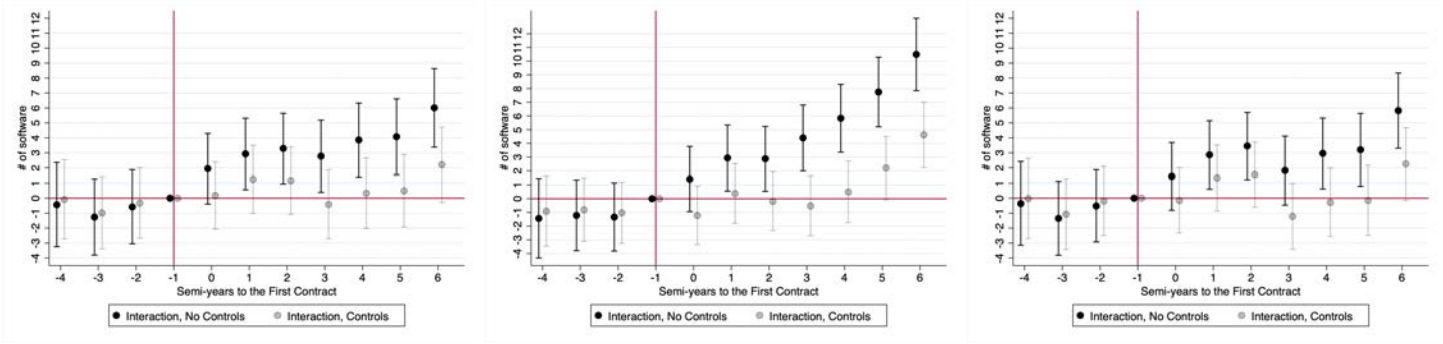


Figure A.9: Differential data-complementary software (left column) and differential AI software development intended for government (middle column) or for commercial uses (right column), resulting from public security contracts that led to government video facial recognition AI software within 1 year, 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: Summary statistics — firms and their production

	Any contract		Public security contract		Public security contract by surveillance capacity	
	Yes	No	Yes	No	High	Low
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Firm characteristics						
Year firm established	2,009.335	2,013.781	2,008.832	2,010.586	2,008.143	2,009.129
	(6.389)	(4.244)	(6.523)	(5.445)	(6.581)	(6.458)
Capitalization (millions USD)	22.964	5.091	39.732	12.682	12.934	61.445
	(210.840)	(43.007)	(333.095)	(149.926)	(35.929)	(464.613)
Rounds of investment funding	2.517	2.046	2.431	2.333	2.500	2.370
	(1.961)	(3.258)	(1.668)	(1.803)	(1.627)	(1.708)
Observations	1,093	6,041	310	497	140	155
Panel B: Software production before contract						
Total amount of software	22.653	14.572	33.368	21.142	33.204	35.579
	(37.860)	(24.473)	(54.760)	(26.759)	(52.535)	(59.246)
Commercial	9.020	6.283	11.821	9.282	12.759	11.404
	(17.087)	(12.502)	(20.580)	(14.007)	(22.587)	(19.503)
Government	7.342	3.951	13.145	5.485	11.407	15.772
	(16.269)	(8.175)	(27.794)	(8.012)	(19.238)	(34.877)
AI-common	3.878	2.580	4.974	3.973	5.500	4.789
	(7.300)	(6.280)	(8.616)	(6.630)	(10.633)	(6.623)
AI-video	1.564	1.026	2.889	1.269	2.741	3.281
	(3.838)	(2.827)	(5.990)	(2.820)	(6.166)	(6.086)
Data-complementary	9.235	5.632	13.821	8.533	12.981	15.140
	(16.704)	(10.763)	(24.485)	(12.831)	(19.440)	(29.441)
Observations	956	6,042	234	443	108	114
Panel C: Software production after contract						
Total amount of software	24.393		35.569	19.903	33.157	40.568
	(59.812)	-	(83.577)	(37.582)	(59.391)	(103.702)
Commercial	8.381		9.968	7.985	11.557	9.226
	(19.628)	-	(18.936)	(14.193)	(21.325)	(17.362)
Government	9.105		16.547	5.664	11.257	22.723
	(34.029)	-	(59.669)	(10.574)	(19.375)	(82.114)
AI-common	4.126		5.601	3.557	5.086	6.503
	(11.623)	-	(15.593)	(7.038)	(9.858)	(19.950)
AI-video	1.584		2.270	1.202	2.286	2.465
	(4.014)	-	(4.860)	(2.941)	(4.524)	(5.344)
Data-complementary	10.046		15.428	7.791	14.957	17.084
	(26.230)	-	(36.886)	(13.484)	(30.123)	(43.561)
Observations	1,095	0	311	411	140	155

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

Table A.3: Summary statistics — procurement contracts

	Non-public security contracts	Public security contracts		
	All	All	Low capacity	High capacity
	(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.

Table A.4: 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.5: Public security contracts vs. non-public security contracts

	Government	Commercial	Data-complementary	Government	Commercial	Data-complementary
	(1)	(2)	(3)	(4)	(5)	(6)
4 Semiyears Before	-0.113	0.007	-0.115	-0.013	0.016	-0.009
	(0.195)	(0.132)	(0.158)	(0.218)	(0.149)	(0.174)
3 Semiyears Before	-0.114	-0.000	-0.074	-0.048	0.006	-0.002
	(0.191)	(0.129)	(0.155)	(0.214)	(0.146)	(0.171)
2 Semiyears Before	-0.103	-0.030	-0.077	-0.075	-0.011	-0.036
	(0.187)	(0.126)	(0.151)	(0.209)	(0.143)	(0.167)
Receiving 1st Contract	0.412**	0.394***	0.450***	0.227	0.346**	0.343**
	(0.186)	(0.126)	(0.151)	(0.208)	(0.142)	(0.166)
1 Semiyear After	0.871***	0.871***	0.971***	0.687***	0.793***	0.837***
	(0.201)	(0.135)	(0.162)	(0.224)	(0.152)	(0.178)
2 Semiyears After	1.465***	1.357***	1.483***	1.286***	1.340***	1.353***
	(0.211)	(0.142)	(0.170)	(0.234)	(0.159)	(0.187)
3 Semiyears After	2.191***	1.983***	2.254***	2.089***	1.915***	2.137***
	(0.223)	(0.150)	(0.180)	(0.248)	(0.169)	(0.198)
4 Semiyears After	2.992***	2.628***	3.195***	3.000***	2.619***	3.107***
	(0.238)	(0.160)	(0.193)	(0.264)	(0.180)	(0.212)
5 Semiyears After	3.881***	3.073***	3.941***	3.937***	3.036***	3.898***
	(0.256)	(0.173)	(0.207)	(0.285)	(0.194)	(0.228)
6 Semiyears After	4.971***	3.761***	4.724***	5.231***	3.832***	4.797***
	(0.278)	(0.187)	(0.225)	(0.309)	(0.210)	(0.247)
4 Semiyears Before × Public Security	-0.149	-0.130	-0.076	-0.158	-0.121	-0.108
	(0.316)	(0.214)	(0.256)	(0.354)	(0.242)	(0.283)
3 Semiyears Before × Public Security	-0.049	-0.133	-0.067	-0.075	-0.128	-0.094
	(0.311)	(0.209)	(0.251)	(0.348)	(0.237)	(0.278)
2 Semiyears Before × Public Security	0.017	0.058	0.012	0.032	0.046	-0.004
	(0.306)	(0.206)	(0.247)	(0.342)	(0.234)	(0.273)
Receiving 1st Contract × Public Security	0.442	0.437**	0.383	0.167	0.254	0.150
	(0.301)	(0.203)	(0.244)	(0.336)	(0.230)	(0.269)
1 Semiyear After × Public Security	0.907***	0.662***	0.728***	0.676*	0.769***	0.505*
	(0.321)	(0.217)	(0.260)	(0.358)	(0.245)	(0.286)
2 Semiyears After × Public Security	1.395***	1.134***	1.261***	1.245***	1.162***	1.045***
	(0.337)	(0.228)	(0.272)	(0.376)	(0.257)	(0.300)
3 Semiyears After × Public Security	1.787***	1.555***	1.760***	1.409***	1.585***	1.484***
	(0.351)	(0.237)	(0.284)	(0.391)	(0.268)	(0.313)
4 Semiyears After × Public Security	2.424***	2.340***	2.511***	2.063***	2.614***	2.335***
	(0.370)	(0.250)	(0.300)	(0.413)	(0.282)	(0.330)
5 Semiyears Before × Public Security	3.012***	3.328***	3.410***	2.524***	3.570***	3.141***
	(0.391)	(0.264)	(0.316)	(0.437)	(0.298)	(0.349)
6 Semiyears After × Public Security	4.068***	4.199***	4.812***	3.568***	4.324***	4.620***
	(0.420)	(0.282)	(0.339)	(0.469)	(0.319)	(0.374)

Notes: Baseline specification (Columns 1–3) controls for time period fixed effects and firm fixed effects. Columns 4-6 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.6: Public security contracts — high vs. low surveillance capacity

	Government	Commercial	Data-complementary	Government	Commercial	Data-complementary
	(1)	(2)	(3)	(4)	(5)	(6)
4 Semiyears Before	-0.177	-0.239	-0.310	-0.182	-0.243	-0.317
	(0.268)	(0.231)	(0.270)	(0.267)	(0.231)	(0.267)
3 Semiyears Before	-0.040	-0.180	-0.118	-0.044	-0.183	-0.123
	(0.264)	(0.228)	(0.266)	(0.262)	(0.227)	(0.262)
2 Semiyears Before	-0.002	-0.202	-0.151	-0.004	-0.203	-0.153
	(0.261)	(0.225)	(0.262)	(0.260)	(0.224)	(0.259)
Receiving 1st Contract	0.750***	0.868***	0.959***	0.680**	0.833***	0.853***
	(0.279)	(0.239)	(0.280)	(0.277)	(0.239)	(0.277)
1 Semiyear After	1.443***	1.663***	1.871***	1.378***	1.630***	1.772***
	(0.289)	(0.250)	(0.291)	(0.288)	(0.250)	(0.288)
2 Semiyears After	2.243***	2.219***	2.576***	2.106***	2.174***	2.367***
	(0.301)	(0.258)	(0.301)	(0.300)	(0.258)	(0.297)
3 Semiyears After	2.986***	3.122***	3.331***	2.917***	3.087***	3.223***
	(0.334)	(0.287)	(0.336)	(0.332)	(0.287)	(0.331)
4 Semiyears After	3.984***	4.017***	4.362***	3.910***	3.980***	4.248***
	(0.360)	(0.309)	(0.362)	(0.358)	(0.308)	(0.357)
5 Semiyears After	4.849***	4.857***	5.662***	4.771***	4.817***	5.543***
	(0.389)	(0.337)	(0.395)	(0.387)	(0.336)	(0.390)
6 Semiyears After	5.595***	5.811***	6.383***	5.511***	5.769***	6.255***
	(0.444)	(0.378)	(0.443)	(0.441)	(0.378)	(0.438)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130	-0.243	0.653	0.176
	(0.620)	(0.539)	(0.627)	(0.617)	(0.538)	(0.620)
3 Semiyears Before × High Capacity	-0.379	0.222	-0.124	-0.356	0.235	-0.099
	(0.565)	(0.488)	(0.570)	(0.562)	(0.487)	(0.563)
2 Semiyears Before × High Capacity	-0.209	0.351	0.118	-0.192	0.362	0.136
	(0.535)	(0.463)	(0.540)	(0.532)	(0.462)	(0.534)
Receiving 1st Contract × High Capacity	0.465	0.314	0.303	0.457	0.307	0.277
	(0.508)	(0.438)	(0.512)	(0.505)	(0.437)	(0.506)
1 Semiyear After × High Capacity	0.858	0.502	0.645	0.817	0.478	0.574
	(0.524)	(0.451)	(0.528)	(0.521)	(0.450)	(0.521)
2 Semiyears After × High Capacity	0.817	0.969**	0.909*	0.814	0.904**	0.890*
	(0.520)	(0.449)	(0.524)	(0.518)	(0.449)	(0.518)
3 Semiyears After × High Capacity	1.023*	0.526	0.963*	0.868	0.442	0.711
	(0.544)	(0.470)	(0.549)	(0.541)	(0.469)	(0.542)
4 Semiyears After × High Capacity	1.151**	0.823*	1.256**	0.987*	0.735	0.988*
	(0.565)	(0.487)	(0.570)	(0.562)	(0.486)	(0.563)
5 Semiyears Before × High Capacity	1.800***	1.205**	1.592***	1.623***	1.110**	1.303**
	(0.594)	(0.515)	(0.602)	(0.591)	(0.514)	(0.595)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***	2.715***	1.759***	2.452***
	(0.642)	(0.550)	(0.644)	(0.638)	(0.549)	(0.636)

Notes: All regressions estimated on the sample of firms with first contracts with a public security agency. Baseline specification (Columns 1–3) controls for time period fixed effects and firm fixed effects. Columns 4-6 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: Robustness

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A.1: LSTM categorization model configuration (vary timestep 10, embeddings 32, nodes 32)			
4 semiyears before	-0.113	-0.310	-0.371
	(0.275)	(0.324)	(0.242)
6 semiyears after	4.637***	4.948***	3.847***
	(0.452)	(0.532)	(0.397)
4 semiyears before × high capacity	-0.328	0.521	0.456
	(0.638)	(0.760)	(0.563)
6 semiyears after × high capacity	2.516***	3.349***	3.579***
	(0.658)	(0.775)	(0.575)
Panel A.2: LSTM categorization model configuration (timestep 20, vary embeddings 16, nodes 32)			
4 semiyears before	-0.268	-0.269	-0.424*
	(0.288)	(0.270)	(0.245)
6 semiyears after	6.102***	4.743***	5.505***
	(0.474)	(0.444)	(0.406)
4 semiyears before × high capacity	-0.181	0.418	0.463
	(0.669)	(0.634)	(0.570)
6 semiyears after × high capacity	2.532***	2.530***	2.513***
	(0.689)	(0.647)	(0.586)
Panel A.3: LSTM categorization model configuration (timestep 20, embeddings 32, vary nodes 16)			
4 semiyears before	-0.206	-0.353	-0.216
	(0.295)	(0.310)	(0.227)
6 semiyears after	6.017***	4.485***	5.667***
	(0.485)	(0.509)	(0.374)
4 semiyears before × high capacity	-0.172	0.526	0.149
	(0.685)	(0.721)	(0.526)
6 semiyears after × high capacity	3.190***	2.652***	2.378***
	(0.706)	(0.741)	(0.541)
Panel B.1: LSTM categorization model threshold (60%)			
4 semiyears before	-0.139	-0.272	-0.309
	(0.234)	(0.309)	(0.255)
6 semiyears after	3.465***	6.452***	5.826***
	(0.389)	(0.508)	(0.421)
4 semiyears before × high capacity	-0.237	0.525	0.553
	(0.543)	(0.721)	(0.595)
6 semiyears aafter × high capacity	2.811***	2.349***	2.765***
	(0.562)	(0.740)	(0.609)
Panel B.2: LSTM categorization model threshold (70%)			
4 semiyears before	-0.133	-0.280	-0.304
	(0.233)	(0.309)	(0.254)
6 semiyears after	3.403***	6.411***	5.789***
	(0.387)	(0.507)	(0.419)
4 semiyears before × high capacity	-0.243	0.542	0.545
	(0.541)	(0.720)	(0.593)
6 semiyears after × high capacity	2.765***	2.324***	2.730***
	(0.560)	(0.739)	(0.607)

Panel C.1: Time frame (full balanced panel)			
4 semiyears before	0.184	0.035	-0.005
	(0.576)	(0.477)	(0.563)
6 semiyears after	5.634***	6.165***	6.614***
	(0.728)	(0.597)	(0.706)
4 semiyears before × high capacity	-3.218	0.743	-0.912
	(2.661)	(2.093)	(2.472)
6 semiyears after × high capacity	3.404***	2.048**	3.071**
	(1.237)	(1.024)	(1.217)
Panel C.2: Time frame (extended time frame)			
5 semiyears before	-0.124	-0.204	-0.245
	(0.274)	(0.236)	(0.275)
8 semiyears after	8.469***	6.986***	7.835***
	(0.572)	(0.488)	(0.562)
5 semiyears before × high capacity	-0.342	0.269	-0.248
	(0.686)	(0.597)	(0.695)
8 semiyears after × high capacity	3.793***	4.150***	5.573***
	(0.756)	(0.648)	(0.750)
Panel D.1: Drop ambiguous public security agencies			
4 semiyears before	-0.184	-0.260	-0.319
	(0.270)	(0.230)	(0.270)
6 semiyears after	5.335***	5.916***	6.094***
	(0.448)	(0.377)	(0.444)
4 semiyears before × high capacity	-0.375	0.625	-0.026
	(0.649)	(0.557)	(0.653)
6 semiyears after × high capacity	3.222***	1.371**	2.897***
	(0.659)	(0.558)	(0.657)

Notes: Specifications include full set of time indicators and interactions with public security 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 A.6, Columns 1-3 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 a timestep of 10 instead; Panel A.2 presents results for the same model trained with an embedding size of 16 instead; Panel A.3 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 A.6, Columns 1-3 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"). Panels B.1 and B.2 replicate the exercise setting the threshold to be higher, at 60% and 70% respectively. Panel C.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 C.2 extends the time frame to 5 semi-years before and 8 semi-years after the initial contracts. Panel D.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%.

Table A.8: 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 × high capacity	-0.267	0.603	0.178
	(0.620)	(0.539)	(0.627)
6 semiyears after × high capacity	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 × high capacity	-0.243	0.653	0.176
	(0.617)	(0.538)	(0.620)
6 semiyears after × high capacity	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 × high capacity	-0.277	0.632	0.131
	(0.620)	(0.539)	(0.627)
6 semiyears after × high capacity	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	-0.167	-0.249	-0.311
	(0.268)	(0.231)	(0.270)
6 semiyears after	5.439***	5.957***	6.404***
	(0.443)	(0.378)	(0.443)
4 semiyears before × high capacity	-0.177	0.526	0.115
	(0.619)	(0.538)	(0.628)
6 semiyears after × high capacity	2.138***	2.605***	2.866***
	(0.645)	(0.553)	(0.648)
Panel A.5: Control for firm age			
4 semiyears before	-0.130	-0.237	-0.282
	(0.263)	(0.231)	(0.269)
6 semiyears after	53.636***	7.926***	28.782***
	(1.226)	(1.078)	(1.261)
4 semiyears before × high capacity	-0.440	0.626	0.050
	(0.608)	(0.539)	(0.625)
6 semiyears after × high capacity	3.279***	1.876***	2.924***
	(0.630)	(0.550)	(0.642)
Panel A.6: All previous controls combined			
4 semiyears before	-0.133	-0.233	-0.326
	(0.262)	(0.230)	(0.265)
6 semiyears after	52.516***	9.002***	28.815***
	(1.224)	(1.078)	(1.250)

4 semiyears before × high capacity	-0.314	0.508	0.126
	(0.605)	(0.537)	(0.617)
6 semiyears after × high capacity	2.688***	1.786***	4.031***
	(0.651)	(0.571)	(0.660)
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 × high capacity	0.170	0.869*	0.489
	(0.538)	(0.514)	(0.586)
6 semiyears after × high capacity	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 × high capacity	0.170	0.841*	0.361
	(0.538)	(0.487)	(0.589)
6 semiyears after × high capacity	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 × high capacity	-0.078	0.869*	0.302
	(0.579)	(0.514)	(0.596)
6 semiyears after × high capacity	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 × high capacity	1.035	1.047	0.820
	(0.786)	(1.384)	(1.081)
6 semiyears after × high capacity	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 × high capacity	-0.114	0.763*	0.235
	(0.634)	(0.404)	(0.603)
6 semiyears after × high capacity	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 × high capacity	-0.053	0.721	0.177
	(0.555)	(0.658)	(0.586)
6 semiyears after × high capacity	2.365***	2.747***	2.815***
	(0.542)	(0.636)	(0.567)

Notes: Specifications include full set of time indicators and interactions with public security 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 A.6 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, Panel A.5 interacts firm age, and Panel A.6 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 A.6. 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*

$$\Pi_c(0, p_c, p_d) = \mu_z \Pi_z(p_z(p_c)). \tag{23}$$

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

$$\kappa_g \frac{\Pi_g(p_g, \bar{d}_g)}{p_g} \frac{\chi}{1+\beta(\chi-1)} = \bar{d}_g. \tag{24}$$

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

$$\frac{q_c(0, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0, p_c, p_d(p_c))} < \frac{Y_c}{D_p} = \frac{\left(\frac{p_c}{1-a}\right)^{-\epsilon}}{\kappa_p} < \frac{q_c(\bar{d}_g, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g, p_c, p_d(p_c))}. \tag{25}$$

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

$$\gamma > 1 + \beta(\chi - 1) \tag{26}$$

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

$$\Pi_g(p_g(\bar{d}_g), \bar{d}_g) + \Pi_c(\bar{d}_g, \bar{p}_c, p_d(\bar{p}_c)) - \left(2 + \frac{F}{\lambda}\right) \Pi_c(0, \bar{p}_c, p_d(\bar{p}_c)) > 0, \tag{28}$$

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:

$$r = \theta\eta + \rho \tag{29}$$

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

$$\tilde{N}_c q_c(0, p_c, p_d)^{1-\frac{1}{\chi}} + N_g q_c(\bar{d}_g, p_c, p_d)^{1-\frac{1}{\chi}} = Y_c = \left(\frac{p_c}{1-a}\right)^{-\epsilon} Y \tag{30}$$

$$N_z q_z(p_z)^{1-\frac{1}{\chi}} = Y_z = \left(\frac{p_z}{a}\right)^{-\epsilon} Y \tag{31}$$

[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 (33). 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.

$$N_g q_g(\bar{d}_g, p_g)^{1-\frac{1}{\chi}} = Y_g = G \tag{32}$$
$$\tilde{N}_c d_p(0, p_c, p_d) + N_g d_p(\bar{d}_g, p_c, p_d) = D_p = \kappa_p Y \tag{33}$$
$$N_g \bar{d}_g = D_g = \kappa_g G, \tag{34}$$

From (4), it is straightforward to see that $p_z(p_c)$ exist and has a negative derivative. Equations (23) 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 (24) 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 (25) 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 (25) to hold.

The optimality condition for private data demand is,

$$d_p{}^{\frac{1}{\gamma}} \left(\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{1}{\gamma-1}\left(\frac{\gamma}{1+\beta(\chi-1)}-1\right)} = \frac{(1-\alpha)}{p_d} (p_c)^{\frac{\chi}{1+\beta(\chi-1)}} \beta \left(\frac{(1-\beta)}{\phi}\right)^{\frac{(1-\beta)(\chi-1)}{1+\beta(\chi-1)}}. \tag{35}$$

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

$$\frac{q_c(\bar{d}_g, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g, p_c, p_d(p_c))} = \frac{\chi}{\chi-1} \frac{p_d(p_c)}{\beta p_c} \left(\frac{\alpha}{(1-\alpha)} \left(\frac{\bar{d}_g}{d_p(\bar{d}_g, p_c, p_d(p_c))}\right)^{\frac{\gamma-1}{\gamma}} + 1\right) \tag{36}$$
$$= \frac{q_c(0, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0, p_c, p_d(p_c))} \left(\frac{\alpha}{(1-\alpha)} \left(\frac{\bar{d}_g}{d_p(\bar{d}_g, p_c, p_d(p_c))}\right)^{\frac{\gamma-1}{\gamma}} + 1\right) \tag{37}$$
$$> \frac{q_c(0, p_c, p_d(p_c))^{1-\frac{1}{\chi}}}{d_p(0, p_c, p_d(p_c))}. \tag{38}$$

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

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

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 (35), 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

We first show the comparative statics of η and n_c with respect to changes in $\bar{d}_g$. We then provide intuition for the result.

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,

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

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

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 (32) and data (34), 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

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

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

$$\frac{dlog(p_d)}{dlog(\bar{d}_g)} = \frac{1}{\beta}\frac{dlog(p_c)}{dlog(\bar{d}_g)} - \frac{1}{\beta}\frac{1}{1+\beta(\chi-1)}\frac{dlog(p_z)}{dlog(\bar{d}_g)}. \tag{39}$$

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 $\epsilon \geq \frac{\chi+\beta(\chi-1)}{1+\beta(\chi-1)}$.

Intuition To understand the theorem, it helps to consider the construction of a BGP equilibrium given an exogenous increase in $\bar{d}_g$ and p_g (instead of just $\bar{d}_g$). 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 (30) and (31), we obtained before:

$$n_c = \underbrace{\left(\frac{1-a}{a}\frac{p_z}{p_c}\right)^{\epsilon}}_{=\frac{Y_c}{Y_z}} \frac{q_z(p_z)^{\frac{\chi-1}{\chi}}}{q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}} + q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}}. \tag{40}$$

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 (34). 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 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 $\bar{d}_g$.

SUPPLEMENTARY MATERIAL (NOT FOR PUBLICATION)

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

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

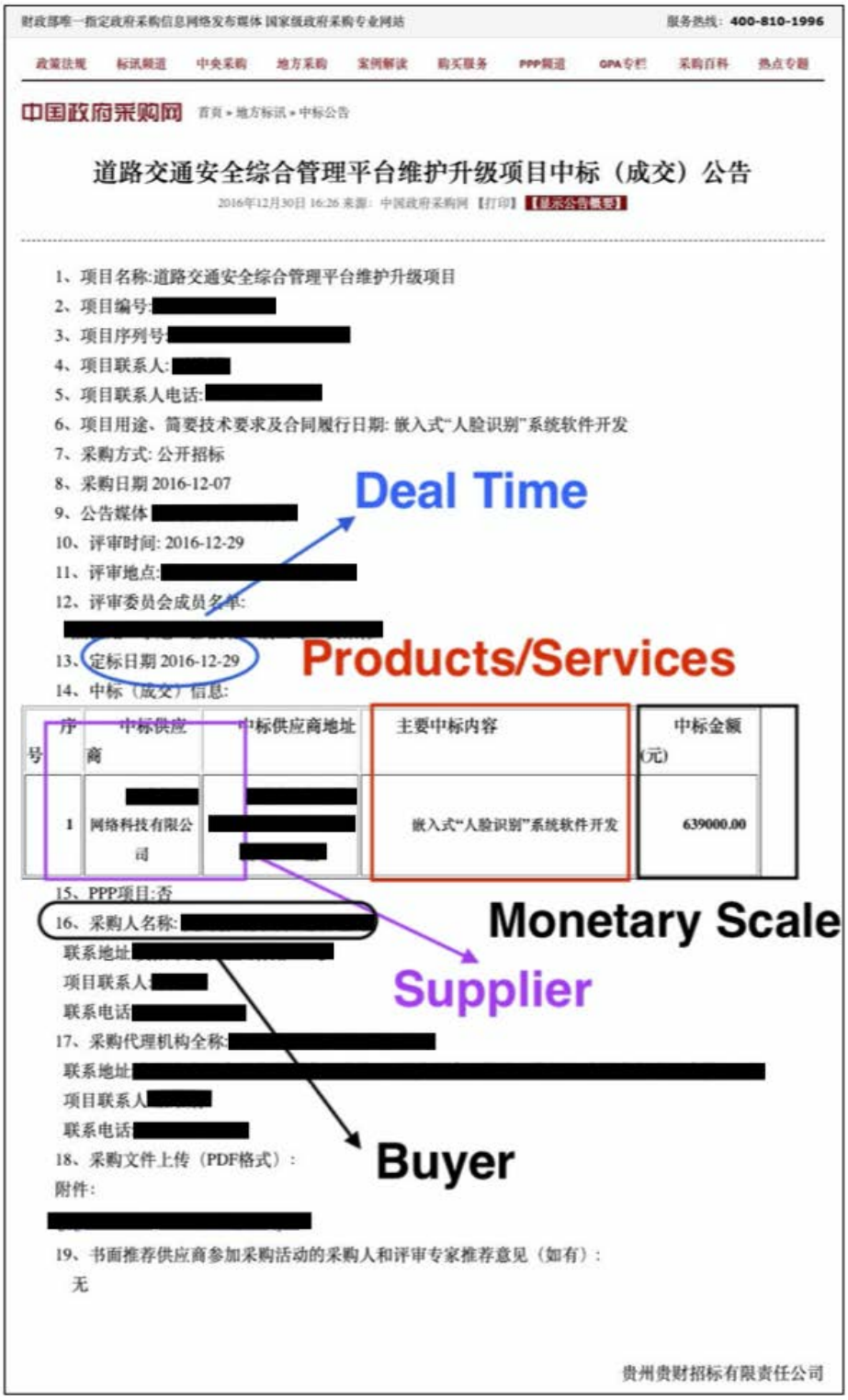
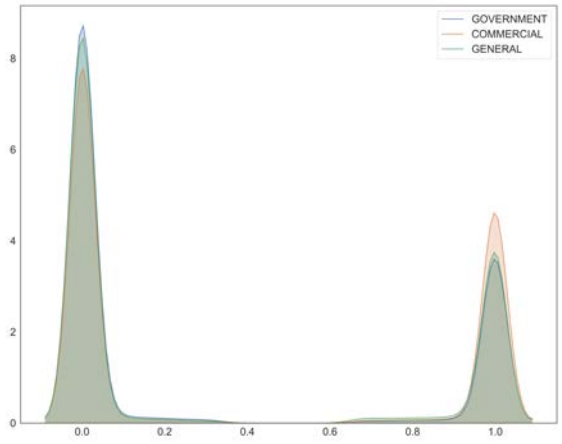
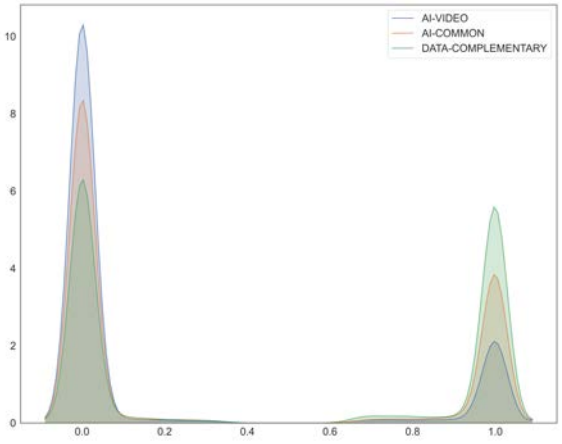


Figure S.3: Example of a procurement contract record; source: Chinese Government Procurement Database.



(a) Customers



(b) Function

Figure S.4: 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.

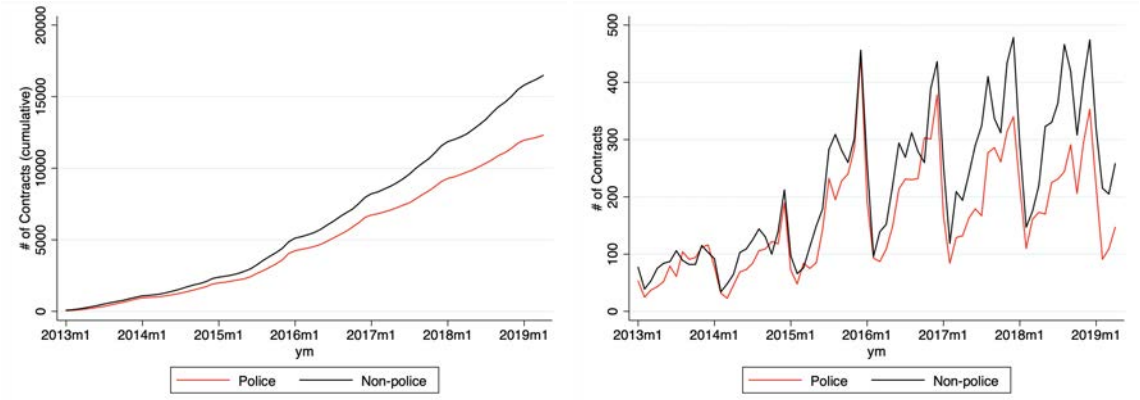


Figure S.5: Cumulative number of public security and non-public security contracts (left panel), and the flow of new contracts signed in each month (right panel).

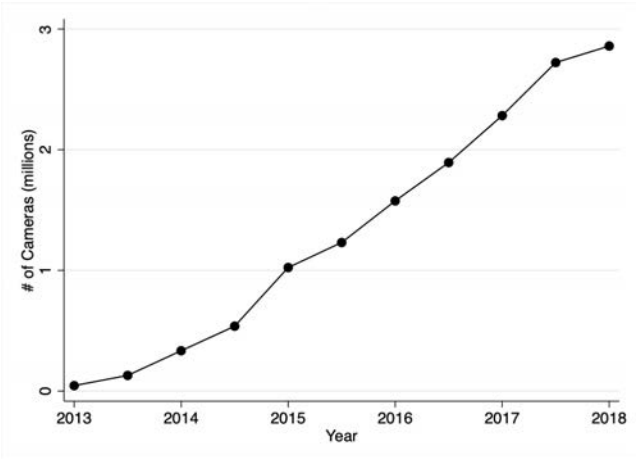


Figure S.6: Number of new public surveillance cameras in China since 2013, as measured by government procurement contracts for cameras. Source: Chinese Government Procurement Database.

	#	Developer	VISA Photos FNMR@ FMR ≤ 0.000001	VISA Photos FNMR@ FMR ≤ 0.0001	MUGSHOT Photos FNMR@ FMR ≤ 0.0001	WILD Photos FNMR@ FMR ≤ 0.00001	CHILD EXP Photos FNMR@ FMR ≤ 0.01	Submission Date
	1	yitu-002	0.004[1]	0.001[1]	0.013[7]	0.052[13]		2018_10_19
	2	yitu-001	0.007[2]	0.003[7]	0.013[8]	0.058[26]	0.579[13]	2018_06_12
	3	sensetime-001	0.009[3]	0.003[6]	0.013[11]	1.000[76]		2018_10_19
	4	sensetime-002	0.010[4]	0.003[10]	0.015[29]	1.000[77]		2018_10_19
	5	siat-002	0.013[5]	0.004[15]	0.014[15]	0.055[20]	0.428[3]	2018_06_13
	6	ntechlab-004	0.013[6]	0.003[4]	0.013[12]	0.046[6]	0.420[2]	2018_06_14
	7	ntechlab-005	0.014[7]	0.002[2]	0.013[10]	0.050[10]		2018_10_19
	8	megvii-002	0.014[8]	0.004[12]	0.030[63]	0.071[35]		2018_10_19
	9	vocord-005	0.016[9]	0.003[3]	0.015[32]	0.048[9]		2018_10_18
	10	everai-001	0.016[10]	0.004[14]	0.013[2]	0.031[2]		2018_10_30

Figure S.7: Face Recognition Vendor Test (FRVT) ranking of top facial recognition algorithms, 2018. Source: *National Institute of Standards and Technology (NIST)*.

Table S.1: 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

Supplementary Material B Quantitative analysis

Supplementary Material B.1 $p_g G/Y$ and D_g/Y as a function of $\bar{d}_g$

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 (36).

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 (35), 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 (39), we obtain

$$\frac{dlog(G/Y)}{dlog(\bar{d}_g)} > -\left((\epsilon+\frac{1-\beta}{\beta})\frac{dlog(p_c)}{dlog(\bar{d}_g)}-\frac{1}{\beta}\frac{1}{1+\beta(\chi-1)}\frac{dlog(p_z)}{dlog(\bar{d}_g)}\right)$$

$$
\begin{aligned}
&- \frac{1}{\gamma}\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p(\bar{d}_g, p_c, p_d))^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + \frac{\gamma}{1+\beta(\chi-1)}(1-\alpha)(d_p(\bar{d}_g, p_c, p_d))^{\frac{\gamma-1}{\gamma}}}\frac{1}{\beta}\frac{\frac{dlog(p_z)}{dlog(\bar{d}_g)} - (1-\beta)\frac{dlog(p_c)}{dlog(\bar{d}_g)}}{1+\beta(\chi-1)} \\
&> -\left((\epsilon + \frac{1-\beta}{\beta})\frac{dlog(p_c)}{dlog(\bar{d}_g)} - \frac{1}{\beta}\frac{1}{1+\beta(\chi-1)}\frac{dlog(p_z)}{dlog(\bar{d}_g)}\right) \\
&- \frac{1}{1+\beta(\chi-1)}\frac{1}{\beta}\frac{\frac{dlog(p_z)}{dlog(\bar{d}_g)} - (1-\beta)\frac{dlog(p_c)}{dlog(\bar{d}_g)}}{1+\beta(\chi-1)} \\
&= -\left(\epsilon + \frac{1-\beta}{\beta}\left(1 - \frac{1}{(1+\beta(\chi-1))^2}\right)\right)\frac{dlog(p_c)}{dlog(\bar{d}_g)} + \frac{1}{\beta}\frac{1}{1+\beta(\chi-1)}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}\frac{dlog(p_z)}{dlog(\bar{d}_g)} \\
&> 0,
\end{aligned}
$$

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 $\bar{d}_g$.

Supplementary Material B.2 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:

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

This implies

$$
\begin{aligned}
\pi_i = p_i &\left(\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\beta\frac{\chi-1}{\chi}} (x)^{(1-\beta)\frac{\chi-1}{\chi}} \frac{1}{\chi-1} \times \cdots \\
&\left(1 + \beta(\chi-1)\frac{\alpha(d_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}}\right) \\
x = &\left(\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}} \left(p_i\frac{1-\beta}{\phi}\right)^{\frac{1}{1-(1-\beta)\frac{\chi-1}{\chi}}} \\
(d_p)^{\frac{1}{\gamma}} = &\frac{(1-\alpha)}{p_d}(p_i)^{\frac{1}{1-(1-\beta)\frac{\chi-1}{\chi}}} \left(\alpha(d_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(d_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}-1} \left(\frac{1-\beta}{\phi}\right)^{\frac{(1-\beta)\frac{\chi-1}{\chi}}{1-(1-\beta)\frac{\chi-1}{\chi}}} \beta,
\end{aligned}
$$

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:

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

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

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

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

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

$$D_p = \left(N_c + \frac{1-\lambda}{\lambda}N_g\right)d_p(0, p_c, p_d) + N_g d_p(\bar{d}_g, p_c, p_d)$$

$$\Pi_z(p_z) = (p_z)^{\frac{\chi}{1+\beta(\chi-1)}}\frac{1+\beta(\chi-1)}{\chi-1}$$
$$Y_z = N_z\frac{\Pi_z(p_z)}{p_z}\frac{\chi}{1+\beta(\chi-1)}.$$

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

$$Y_z = \left(\frac{p_z}{a}\right)^{-\epsilon} Y$$

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

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.

Supplementary Material B.3 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 A.6, 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 A.6, 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 $\alpha = 0.82$.