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

Martin Beraja
David Y. Yang
Noam Yuchtman*

July 17, 2020

Abstract

Data-intensive technologies, like AI, are increasingly widespread. We argue that states may play an important role in shaping data-intensive innovation because: *(i)* states are key collectors of data and *(ii)* data is sharable across uses within firms, potentially generating economies of scope. We investigate a prototypical setting: facial recognition AI in China. Collecting comprehensive data on AI firms and government procurement contracts, we find economies of scope from government data: firms awarded contracts richer in government data produce *both* more government and commercial software. To study the aggregate implications of government data provision, we build a directed technical change model. We show that government data provision can increase growth and bias the direction of innovation, but welfare increases only when economies of scope are sufficiently strong and citizens' and states' preferences are sufficiently aligned. We conclude with three applications analyzing states' choices of: industrial policy, surveillance levels, and privacy regulation.

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, Ernesto Dal Bó, Dave Donaldson, Ruben Enikolopov, Richard Freeman, Andy Neumeyer, Juan Pablo Nicolini, Arianna Ornaghi, Maria Petrova, Torsten Persson, Nancy Qian, John Van Reenen, Andrei Shleifer 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; to surveillance data used to provide public security; to geographic and scientific data used for national defense, 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 for commercial ones. Thus, government data collection and provision can potentially affect not only innovation of government products, but also innovation targeting much larger commercial markets. In this paper, we argue that, because of these two features of data, states may play an important role in shaping innovation and growth in data-intensive economies.

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 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 provide 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 government provision of data, we build a general equilibrium directed technical change model where some firms choose to engage in data-intensive innovation 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 in-

[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 ..." (`https://asiahouse.org/news-and-views/kai-fu-lee-age-ai-china-new-opec/`, last accessed July 10, 2020).

[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 in the presence of sharable inputs.

novation crowds-out resources from consumption, government data provision increases welfare only when economies of scope are sufficiently strong.[3] The welfare implications of government data collection and provision are further complicated by their potentially harmful social and political consequences. We conclude with three applications which illustrate the varied roles that states may play in data-intensive economies: first, in setting industrial policy; second, in choosing the level of public surveillance; and third, in enacting privacy regulation.

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 provides an exogenous increase in the government data available to a firm will lead to increased production of *both* government and commercial software. In practice though, 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, in particular allowing linking faces to personal identifiers. The Chinese state both collect huge amounts of personally identifiable data and demand facial recognition software for surveillance purposes. A firm receiving a government contract would thus receive access to government data which is not publicly available, using this data to develop the software it was contracted to produce. For example, when obtaining a contract with a municipal police department to produce surveillance software, it would receive access to video from street cameras and a database of labeled personal images. It would then develop surveillance software by training an AI algorithm that matches individuals in video to the database of labeled images. Crucially though, 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.[4] 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

[3]These results, albeit different in context and mechanism, echo those in Barro (1990), who shows how optimal government provision of services (like infrastructure) trades-off direct increases in firm productivity and growth against the crowding-out of resources from consumption.

[4]While the ultimate behaviors predicted for government and commercial purposes are likely to differ, the detection problem — and thus the benefit from access to larger amounts of government data — is shared.

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.[5] 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 possessing 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 the second 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

[5]Non-public security agencies do not have access to large scale surveillance camera networks and cover narrower groups of individuals. For example, a bank, school, or hospital might hire an AI firm to provide facial recognition-based access to its facilities. We thus view contracts with non-public security agencies as providing firms with other potential inputs supporting innovation, but *not* access to massive datasets.

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 — a type of software that 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 as well.

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 show that learning-by-doing is unlikely to be differentially stronger for data-rich contracts: data-rich contracts do not meaningfully differ from data-scarce ones in their content and we find that the types of government software produced after data-rich contracts are not different from those produced after data-scarce ones. Moreover, we show that controlling for pre-contract software production (a measure of how strong the potential for learning is) only slightly reduces the differential effects of a data-rich contract. Third, we show that the baseline results are not driven by the differential capital, signaling value, commercial opportunities, or connections to local government that may be provided by data-rich contracts, relative the data-scarce ones.

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, 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 instead used to produce "surveillance services" which are purchased by the state. 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. Finally, a representative household owns all firms and consumes the final good.

We show conditions under which there is a unique 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 provision affects innovation and growth. When commercial software and non-software are weak substitutes, an increase in government data provided to firms increases the BGP rate of innovation and biases private innovation towards 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 of 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 supply it and provide the government data that is produced as a by-product to government software producers. 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.

This macroeconomic framework allows us to illustrate the varied roles that states may play in shaping data-intensive economies. In three applications, we demonstrate that because of the features of data that we highlight: *(i)* industrial policy in the form of government data provision can be justified on grounds which differ from those that motivate traditional industrial policy; *(ii)* surveillance states' desire to monitor and control their citizens aligns with promoting data-intensive innovation but may be detrimental to citizens; and, *(iii)* regulation limiting government data collection due to privacy concerns reduces

innovation but may benefit citizens overall. When we consider these roles of the state — and incorporate into our model the possible divergence of citizens' and states' preferences for data collection — the welfare implications of government data collection and provision become more ambiguous. Surveillance states' data collection might increase growth rates and promote data-intensive innovation, but reduce citizen welfare due to violations of civil liberties. States regulating data collection reflecting citizens' privacy concerns might exhibit less innovation, growth (and consumption), but might make citizens better off.

In what follows, we discuss the related literature and our contribution in Section 2. In Section 3, we present the simple conceptual framework on economies of scope arising from government data as an input, which provides testable predictions that guide our empirical exercise. Next, we present the empirical exercise examining the role of access to government data in shaping innovation in China's facial recognition AI sector: first the empirical context and data sources in Section 4; then the empirical strategy and results in Section 5. In Section 6, we introduce a general equilibrium framework of directed technical change to study the macroeconomic implications of government data provision. We discuss three applications of the framework in Section 7, and offer concluding thoughts in Section 8.

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). We contribute to this literature by examining the role of the state in shaping economic outcomes in the age of data-intensive innovation, 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); Simon and Sichelman (2017) studies the effects of data arising from innovations; 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 both the literature on 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.[6] Recent research on

[6]Contexts in which industrial policy was shown to be effective include: the 19th century French textile

innovation policy also suggests an important role for the state in encouraging R&D — see Bloom et al. (2019).[7] 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.[8] Third, we provide a justification for government data provision that differs from that of traditional industrial policies. For example, Costinot et al. (2019) make the case for industrial policy to correct for learning-by-doing externalities. We show that because states are key collectors of data, and because government data can give rise to economies of scope, it may be optimal to directly provide such data to data-intensive software producers. In this sense, we also contribute to a macroeconomic literature on the role of government spending in promoting economic growth (e.g., Rosenstein-Rodan, 1961, 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 (Hicks, 1932; Habakkuk, 1962; Acemoglu, 2002, 2007). The theoretical literature is well developed, including applications to: climate change and environmental policy (Acemoglu et al., 2012; Hemous, 2016), technology-skill complementarity (Acemoglu, 1998), the sources of cross-country productivity differences (Acemoglu et al., 2006b), migration (Lewis, 2013), and pharmaceutical innovation (Acemoglu and Linn, 2004). We add to this literature by studying a novel (and important)

industry, protected by the blockade of British competitors during the Napoleonic Wars (Juhàsz, 2018); 19th century UK and Great Lakes US shipbuilding (Hanlon, 2020); post-WWII Finland following industrialization imposed by the Soviet Union (Mitrunen, 2019); post-WWII US as a result of Office of Scientific Research and Development (OSRD) spending (Gross and Sampat, 2020); post-WWII Italy, as a result of the US Marshall Plan (Giorcelli, 2019); East Asia's (and China's) Growth Miracle (Lane, 2017; Liu, 2019); and, Chinese shipbuilding in the 2000s (Kalouptsidi, 2017; Barwick et al., 2019). Bartelme et al. (2019) estimate the importance of sectoral economies of scale that are often used to justify industrial policy, finding that industrial policy may not be as effective as other policies (e.g., trade).

[7] Among others, Bronzini and Iachini (2014) find that R&D subsidies by the Italian government raise investment of small enterprises; Howell (2017) shows that the US Department of Energy's funding helps financially constrained firms to attract future funding and innovate; Azoulay et al. (2018) demonstrate that public scientific grant funding increases private sector patenting among pharmaceutical and biotechnology firms; and, Moretti et al. (2019) show that defense-related R&D expenditures of OECD countries crowd in private sector R&D spending. In related work, Moser (2005) and Moser and Voena (2012) study the role of state policy on intellectual property rights in shaping patterns of innovation.

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

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. Existing work has focused on the effect of energy prices on innovation in energy-saving technologies (Newell et al., 1999; Popp, 2002; Aghion et al., 2016), and the effect of demand forces on pharmaceutical innovation (Acemoglu et al., 2006a; Acemoglu and Linn, 2004; Costinot et al., 2019). A notable exception is Hanlon (2015), who documents that the reduction in the supply of American cotton to Britain due to the US Civil War induced British innovation toward technologies that complemented cotton varieties from other sources. 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.[9] In the domain of AI technology, however, surveillance states' objectives and data collection, along with the economies of scope arising from data as an input, facilitate data-intensive innovation even for commercial applications. Thus, the alignment between the state and private sector could offset the expropriation risks and commitment problems traditionally faced by private entrepreneurs under autocracy, although, as we emphasize, such alignment may still be detrimental to citizens overall.[10] 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.[11]

[9]The risk of *ex post* taxation or expropriation of entrepreneurs will mean *ex ante* less investment (North, 1991; 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, 1993, 2002).

[10]Interestingly, placing data in private hands (perhaps closer to the outcome in democratic societies) may produce incentive misalignment *greater* than that arising from political elites' attempts to protect their political rents. In the private sector, an incumbent firm that possesses a large amount of data may not share that data with competitor firms (despite economies of scope arising from data) in order to preserve its *economic* rents.

[11]A large literature studies the Chinese economy and its spectacular growth in the recent decades (e.g., Song et al., 2011; Khandelwal et al., 2013; Roberts et al., 2017; Cheng et al., 2019; Bai et al., 2019), as well as innovation in China more specifically (e.g., Wei et al., 2016; Bombardini et al., 2018). Much of the work on China's political economy highlights institutional features — for example, competition for promotion (e.g.,

3 Economies of scope from government data as an input

This section discusses how government data can generate economies of scope and derives the key testable implication that guides our empirical approach in the following sections. We consider a setting that incorporates the two characteristics of data that we highlight: *(i)* the state is a key collector of data with no perfect private substitutes, and *(ii)* data can be shared across uses.

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 which can be linked to identifiable records) and that are not made publicly available. 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 software for the government 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 their shared uses, consider video from street surveillance cameras and administrative records with the names of individuals linked to images of their faces. This data is used to train an algorithm with the ability to *recognize* faces in video and identify individuals in administrative records. That trained identification algorithm may then also be part of a more complex software application that performs the *predictive* task of identifying potential security threats. That same data, though, is also a crucial input to train algorithms that perform a wide range of *commercial* recognition and prediction tasks, such as identifying a customer in video from cameras in a store or predicting a customer's purchases.[12]

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

Li and Zhou, 2005; Jia et al., 2015), bureaucratic rules of evaluation and rotation (Li, 2019), or social norms (Tsai, 2007) — that allow China to grow *despite* the lack of institutional constraints on the Chinese Communist Party. In our contrast, our work (along with others, like Bai et al., 2019) identifies a mechanism through which autocratic power can actually *promote* economic growth.

[12]Note that an alternative plausible specification of the technologies would be 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 specifications are equivalent.

of the other inputs, and thus at lower cost.[13] 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.[14] 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 both more government and commercial software.

4 The state and China's facial recognition AI: context and data sources

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

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

[14] It is also possible that some of the other inputs, like non-government data, may be substitutable for some types of government data. This would further diminish the effects of a data-rich contract on commercial software production.

development of the technology requires access to large datasets that allow linking faces to personal identifiers. Second, because the Chinese state collects huge amounts of personally identifiable data and demands 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."[15] 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."[16]

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 includes video from street surveillance cameras as well as labeled images with names and faces of individuals. The firm uses this data to train a "detection" AI algorithm: matching faces observed in cameras 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 detect and track individual shoppers throughout their stores, and then predict their consumption choices.

This context allows us to empirically test for economies of scope arising from access to government data, guided by the predictions of our model. 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

[15] Source: Chinese People's Political Consultative Conference official website: `http://www.rmzxb.com.cn/c/2019-06-13/2363368.shtml`, last accessed June 12, 2020.

[16] *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 more broadly a characteristic of AI innovation around the world.

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.[17] We discuss our approach to resolving the data challenges we face, in turn.

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 the People's Bank of China (i.e., China's central bank; see Appendix Figure A.1 for an example). 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 (see Appendix Figure A.2 for an example). Using the overlap between sources, we validate the coding of firms identified in the *Tianyancha* database. We also supplement the *Tianyancha* data by adding a small number of AI firms that are listed by *Pitchbook* but omitted by *Tianyancha*. Overall, we identify 7,837 Chinese facial recognition AI firms.[18] We also collect an array of firm level characteristics such as founding year, capitalization, major external financing sources (such as venture capital rounds), 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 (see Appendix Figure A.3 for an example of such contract). 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

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

[18] These firms fall into 3 categories: *(i)* firms specialized in facial recognition AI (e.g., Yitu); *(ii)* hardware firms that devote substantial resources to develop AI software (e.g., Hik-Vision); and *(iii)* facial recognition AI units of large tech conglomerates (e.g., Baidu AI).

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.[19] 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. We are thus able to distinguish 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").[20] We also distinguish between 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") and those that are data-complementary, involving data storage, data transmission, or data management (e.g., "a computer cluster for webcam monitoring data storage").[21]

To implement the 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 Tables A.3 and A.4 present the top words (in terms of frequency) used for the categorization; Appendix Figure A.4 presents the density plots of the algorithm's category predictions; Appendix Figure A.5 shows the summary statistics of the categorization output by customers and

[19]The prospectus was filed with the Hong Kong Stock Exchange. See `https://go.aws/37GbAZG`, last accessed June 22, 2020.

[20]By coding as "commercial" only those products that are explicitly and specifically linked to commercial applications, and excluding products with ambiguous use, we aim to be conservative in our measure of commercial software products.

[21]We also separately identify within the category of AI software a subcategory in which innovation 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. The differences in 1-to-1, N-to-1, and N-to-N matching in facial recognition are sometimes referred to as facial authentication versus facial recognition; see Gates (2011) for more details.

by function; and finally, Appendix Figure A.6 presents the confusion matrix (Type-I and Type-II errors) of the predictions relative to categorization performed by humans.[22]

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 specifies 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 the public security ones.

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.[23]

Our measure of public security contracts is comprehensive: we capture the following four types of contracts from the Chinese Government Procurement Database procurement contract database: *(i)* all procurement contracts for China's flagship surveillance/monitoring projects — *Skynet Project*, *Peaceful City Project*, and *Bright Transparency Project*; *(ii)* all procurement contracts with local police departments, including traffic police and highway administration bureaus; *(iii)* all procurement contracts with the border control and national security units; and, *(iv)* all procurement contracts with the administra-

[22]The algorithm is very accurate in categorizing software and patents for government purposes. The algorithm is relatively conservative in categorizing software products for commercial customers, and relatively aggressive in categorizing them as for general purpose. In setting our categorization threshold for commercial software we again aim to be conservative in our measure of commercial software products.

[23]We identify 410,510 public security contracts in total. We present the cumulative number of procurement contracts of each type (public security and non-public security) in Appendix Figure A.7, top panel; as well as the flow of new contracts signed in each month (bottom panel). Both the public security and non-public security contracts have steadily increased since 2013.

tive 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; overall, 7.2% (568 of 7,837) of the facial recognition AI firms in our dataset receive at least one public security procurement contract. We find that 12.6% (984 of 7,837) of the facial recognition AI firms in our dataset receive at least one non-public security contract, and 5.2% (408 of 7,837) 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. These are contracts with public security agencies possessing greater video surveillance capacity, which we measure using 5,837 prefectural government contracts for surveillance cameras.[24] 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.[25] 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.[26] 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: we are able to compare outcomes within a set of firms that selected into a similar set of public security contracts.

In Table 1, we present summary statistics describing the firms in our sample. We begin by providing summary statistics splitting firms into those that receive a government con-

[24]These contracts contain data including the quantity of cameras ordered, the total size of the contract, the unit price of cameras, as well as the locality and time in which the contract must be fulfilled; when data on price and quantity is (occasionally) missing, we use data from the same prefecture to impute values. There are on average 77 contracts per prefecture. In Appendix Figure A.8, we present a time series plot of the number of cameras in our data over time.

[25]This measure captures the *stock* of recent surveillance camera purchases at the time an AI procurement contract was awarded. While we do not observe the entire stock of surveillance cameras, we believe that a focus on newer cameras is appropriate given their higher resolution and thus greater usefulness in identifying and matching faces. This is affirmed in the government's directive "Several Opinions on Strengthening the Construction, Networking and Application of Public Security Video Surveillance", issued in 2015 jointly by the Central Committee for Comprehensive Management of Public Security, Ministry of Science and Technology, Ministry of Industry and Information Technology, and Ministry of Public Security, which set camera upgrading as a main goal for China's public security agenda. Source: `https://bit.ly/3dqdjU0`, last accessed on June 22, 2020.

[26]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.

Table 1: Summary statistics — firms and their productions

	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	2,010.052	2,010.547
	(6.389)	(4.244)	(6.523)	(5.445)	(5.806)	(5.979)
Capitalization (millions USD)	22.964	5.091	39.732	12.682	6.597	21.245
	(210.840)	(43.007)	(333.095)	(149.926)	(13.880)	(131.463)
Rounds of investment funding	2.517	2.046	2.431	2.333	2.467	2.538
	(1.961)	(3.258)	(1.668)	(1.803)	(1.613)	(1.890)
Panel B: Software production before contract						
Total amount of software	22.653	14.572	33.368	21.142	34.630	33.100
	(37.860)	(24.473)	(54.760)	(26.759)	(58.228)	(63.319)
Commercial	27.492	28.415	33.709	30.745	33.726	39.050
	(100.363)	(198.404)	(75.537)	(132.801)	(55.705)	(106.153)
Government	20.132	13.810	33.359	18.023	25.726	37.775
	(60.669)	(76.379)	(66.071)	(71.590)	(40.810)	(81.822)
AI-common	12.673	11.541	14.222	14.302	10.877	17.375
	(34.360)	(65.480)	(30.832)	(40.768)	(16.898)	(41.654)
AI-video	7.669	6.071	13.889	6.271	12.822	17.050
	(19.175)	(32.115)	(31.492)	(12.481)	(33.010)	(35.157)
Data-complementary	26.666	21.788	39.385	25.454	33.945	42.250
	(80.688)	(125.330)	(82.107)	(95.536)	(53.336)	(104.592)
Observations	956	6,042	234	443	73	80
Panel C: Software production after contract						
Total amount of software	24.393		35.569	19.903	23.062	24.410
	(59.812)	-	(83.577)	(37.582)	(46.597)	(74.627)
Commercial	47.475		66.000	34.360	20.146	24.889
	(304.587)	-	(504.701)	(125.475)	(44.433)	(93.726)
Government	29.459		43.518	19.762	15.917	28.513
	(120.235)	-	(166.724)	(65.190)	(29.331)	(105.435)
AI-common	22.384		25.363	17.484	8.833	12.308
	(107.543)	-	(116.453)	(81.909)	(20.766)	(35.742)
AI-video	14.422		20.958	8.015	8.104	16.615
	(66.104)	-	(88.789)	(23.440)	(18.484)	(64.953)
Data-complementary	45.489		63.402	28.618	22.000	30.171
	(212.602)	-	(283.432)	(106.427)	(51.466)	(99.271)
Observations	1,095	0	311	411	96	117

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

tract (column 1) and those that do not (column 2). Next, we split firms receiving a contract into those with a first contract that is with a public security agency (column 3) and those with a first contract that is with a non-public security agency (column 4). Finally, we split firms with a first contract with a public security agency into those with a first contract that is data rich (column 5) and those with a first contract that is data scarce (column 6).

One can see that firms receiving different types of contracts differ substantially from each other, so accounting for differences (both observable and unobservable) between the

Table 2: 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.

firms receiving data-rich and data-scarce contracts will be crucial to identify the effects of the contracts. Interestingly, patterns of selection into contracts that are data-rich differs depending on the definition used: for example, firms receiving public security contracts are better capitalized than firms receiving non-public security contracts, but firms receiving public security contracts in high-surveillance prefectures are less well capitalized than firms receiving public security contracts in low-surveillance prefectures. This suggests that very simple selection stories will not easily account for effects of data-rich contracts seen along both margins of comparison.

In Table 2, we present summary statistics describing the contracts procuring AI services in our sample. We begin, in Panel A by describing all of the AI contracts, and then in Panel B, presenting analogous descriptive statistics but limited to the set of firms' *first* contracts.[27] We present contract characteristics, characteristics of the government agen-

[27]In Appendix Table A.2, we provide descriptive statistics for the prefectures where contracts were issued, again disaggregating by the type of agency and by surveillance capacity.

cies awarding the contracts, and characteristics of the municipalities where contracts were awarded, disaggregating the contracts by the issuing agency and by the video surveillance capacity of the prefecture where the contract was issued. In column 1, we present summary statistics for the non-public security contracts; in column 2, we present summary statistics for the full set of public security contracts. In columns 3 and 4, we split the set of public security contracts into data-scarce public security contracts (surveillance capacity below median) and data-rich public security contracts (surveillance capacity above median), respectively.

One can see that 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, 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. Finding consistent effects of data-rich contracts across definitions will argue against simple alternative hypotheses regarding unobserved contract characteristics.

5 Analyzing the role of government data in Chinese facial recognition AI

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 innovation. 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.[28]

The coefficients on the interaction terms (i.e., on $\sum T_{it} \times Data_i$) non-parametrically capture a firm's differential production of new software (or patents) 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 and innovation common to all facial recognition firms (for example, government industrial policy promoting AI), we include time period fixed effects, α_t in all specifications. We also include firm fixed effects, γ_i, in all specifications, allowing us to control for all (observable or unobservable) time-invariant firm characteristics. Finally, in addition to estimating a parsimonious model without controls, we also estimate a model including a vector of pre-contract firm characteristics (X_i) interacted with time period fixed effects.[29] We allow the error term ϵ_{it} to be correlated not only across observations for a single firm, but also across observations for firms that are related by common ownership by a single mother firm.[30]

Our empirical strategy allows us to address important threats to identification. A particular concern is non-random assignment of contracts to firms. 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

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

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

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

and a prefecture's GDP per capita interacted with time period fixed effects. In addition to including this wide range of controls, we will also present more direct evidence of data's importance in generating the effects we observe, as well as additional evidence against alternative mechanisms.

5.2 Results

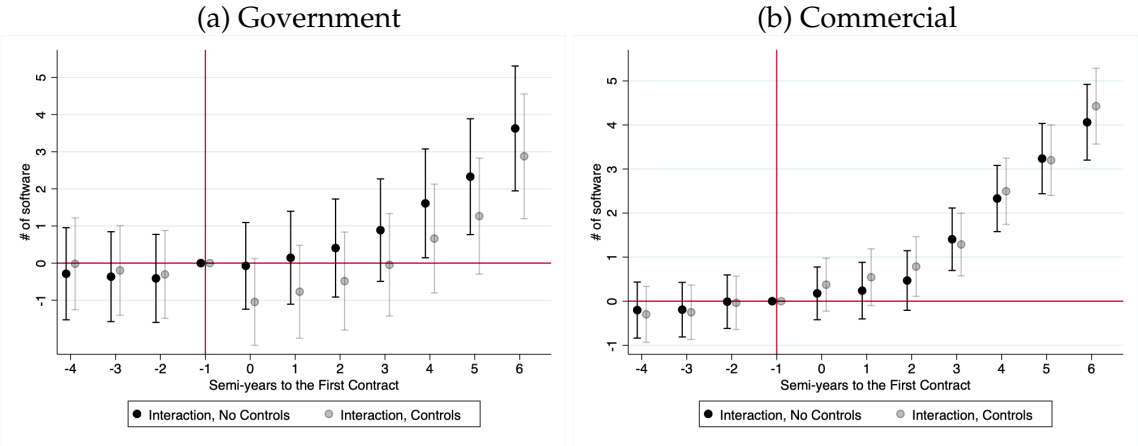
We first present the baseline results in Section 5.2.1 following the empirical strategy described above; we then present evidence suggesting an important role of government data in firms' innovation in Section 5.2.2; finally, in Section 5.2.3 we discuss alternative hypotheses and present evidence suggesting they are not driving our main results.

5.2.1 Baseline estimates and robustness checks

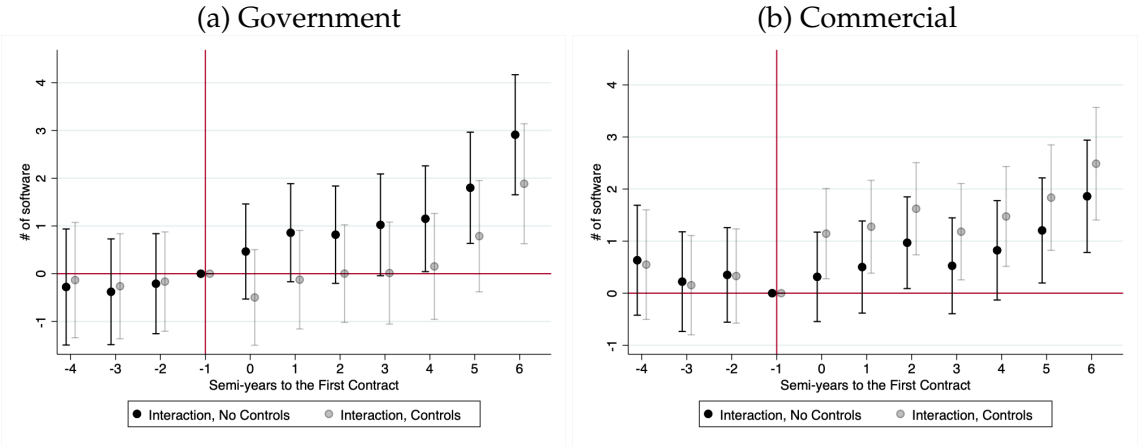
We first estimate our baseline specification, comparing the effects of public security contracts to non-public security contracts on firms' production of software. In Figure 1, Panel A, we plot the coefficients β_{1T}, describing the *differential* software production around the time when a public security 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, and 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 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.[31]

We next, in Figure 1, 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.

[31] One may wonder what are the overall effects of government contracts that underly the differential effects in Figure 1, Panel A. In Appendix Figure A.9, we plot the coefficients β_{2T}, describing software production around the time when a non-public security contract was received, and the sum of the coefficients, $\beta_{1T} + \beta_{2T}$, describing software production around the time when a public security contract was received. We find that government software and commercial software *both* significantly increase after receipt of *both* non-public security and public security contracts, with effects being significantly greater in the latter, as seen in Figure 1, Panel A.



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



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

Figure 1: 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, and presents their effects relative to non-public security contracts. Panel B defines data-rich contracts as public security contracts in prefectures with above-median surveillance capacity, and presents their effects relative to public security contracts in prefectures with below-median surveillance capacity. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.

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 1, 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 has little effect on our findings. In Figure 1, Panel B(b), one again 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 1, 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.[32]

Interpretation As stated 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 sources of data (or other inputs, more generally) 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 government contracts on commercial software production, and differentially large effects of data-rich contracts (see Appendix Figure A.10).

Our findings of economies of scope are all estimated using quantities of software.

[32] We show the overall effects of data-rich and data-scarce public security contracts in Appendix Figure A.10. We plot the coefficients β_{2T}, describing software production around the time when a data-scarce public security contract was received, and the sum of the coefficients, $\beta_{1T} + \beta_{2T}$, describing software production around the time when a data-rich public security contract was received. We again find that government software and commercial software *both* significantly increase after receipt of *both* data-scarce and data-rich contracts, with effects being significantly greater in the latter, as seen in Figure 1, Panel B.

While we cannot observe differences in the quality of software produced as a result of receiving access to more government data, we do observe large increases in the most data-intensive form of facial recognition AI: that using video (see Appendix Figure A.11).

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 (non-public security contracts or public security contracts in prefectures with below-median surveillance capacity), we account for (time-varying) selection into receiving a data-rich 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. First, we can vary the construction of the explanatory variable of interest, adjusting our classification of (data-rich) public security contracts to exclude any ambiguous government agencies (e.g., contracts with the government headquarters, and smart city management and administrative bureaux could be meant to provide security services just for the government office building). This has no impact on our results (see Appendix Table A.7). Next, we assess the robustness of our results to the three key parameters of choice in the RNN algorithm that we use to categorize software — timestep, embedding, and nodes. We vary these three parameters, re-configure the RNN LSTM algorithm, re-categorize software, and re-estimate the baseline empirical specification. We find that these algorithm parameter choices have no impact on our results (see Appendix Table A.8). In addition, we evaluate the robustness of our results to adjustments of the LSTM classification threshold — the baseline specification sets the threshold as 50%. We re-categorize software using higher classification thresholds of 60% and 70%, and these adjustments have no impact on our results (see Appendix Table A.9. Finally, we can vary the time-frame studied: we examine wider windows of time around the receipt of the first contract; and, we consider a balanced panel of firms within a narrow window (studying a balanced panel over too long a window substantially reduces the sample size). These changes, too, have no impact

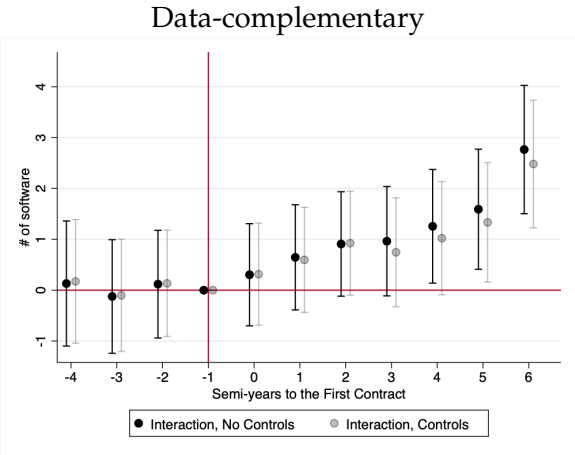


Figure 2: Differential data-complementary software development resulting from data-rich public security contracts, relative to data-scarce public security contracts, controlling for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.

on our findings (see Appendix Table A.10).

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 among which this information is available to estimate the relationship between bid values and local surveillance camera capacity at the time the contract was awarded, as well as the relationship between the number of bidders and local surveillance capacity. The patterns match what we expect (see Appendix Figure A.12): data-rich contracts are associated with lower bids — even controlling for bidding firm fixed effects (p-value $= 0.13$) — and with significantly 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 and transmission). We next test whether firms receiving data-rich contracts differentially produce data-complementary software. Importantly, these data-complementary software products are *distinct* from the *AI* software studied above. In Figure 2, we present estimates from the same specification as in Panel B of Figure 1, 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 con-

tract.[33] 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.[34]

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.[35] We examine whether these firms also differentially produce more government and commercial software after receiving a data-rich public security contract. One can see in Appendix Figure A.15 that indeed they do. Moreover, 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 wide range of tests, exploiting multiple margins of variation in access to government data, all point in the same direction: contracts that provide more government data allow firms to produce more government and commercial software. Beyond other mechanisms through which contracts may affect firm 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 parsimonious presentation of the varied empirical exercises to come, we plot regression coefficients and confidence intervals only for differential effects of data-rich contracts 3

[33] We plot the coefficients β_{2T}, describing data-complementary software production around the time when a data-scarce public security contract was received, and the sum of the coefficients, $\beta_{1T} + \beta_{2T}$, describing data-complementary software production around the time when a data-rich public security contract was received, in Appendix Figure A.13. 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, as seen in Figure 2, Panel A.

[34] 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.14, one can see that public security contracts that led to data-complementary software production within 1st year of the contracts were associated with differentially more government *and* commercial software production. Again we find no evidence of pre-contract differences in software production levels or trends.

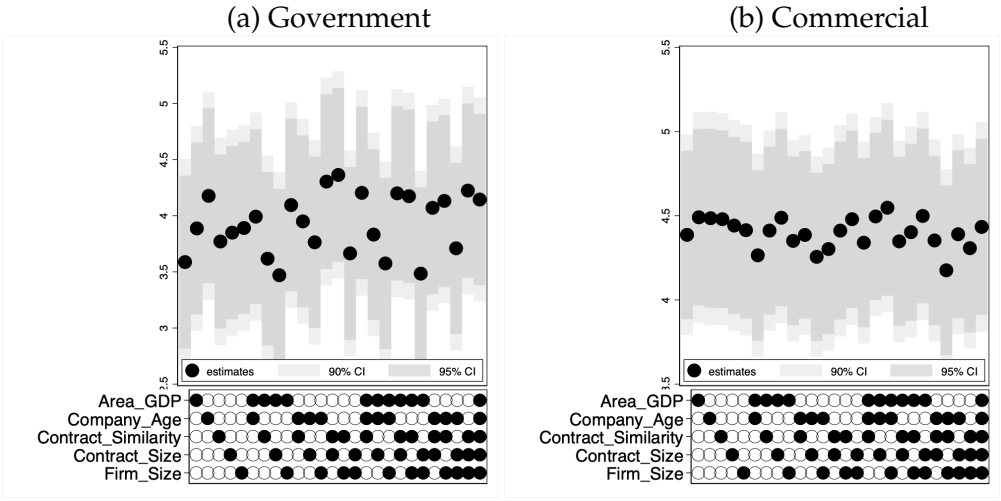
[35] Firms that produce video facial recognition AI for the government after receiving a data-rich public security contract also differentially produce more data-complementary software post-contract. See Appendix Figure A.15.

years following contract receipt, in Figure 3. The figure plots these estimates specification-by-specification for a wide range of specifications. We also present more complete sets of coefficient estimates in tables provided in the Appendix.

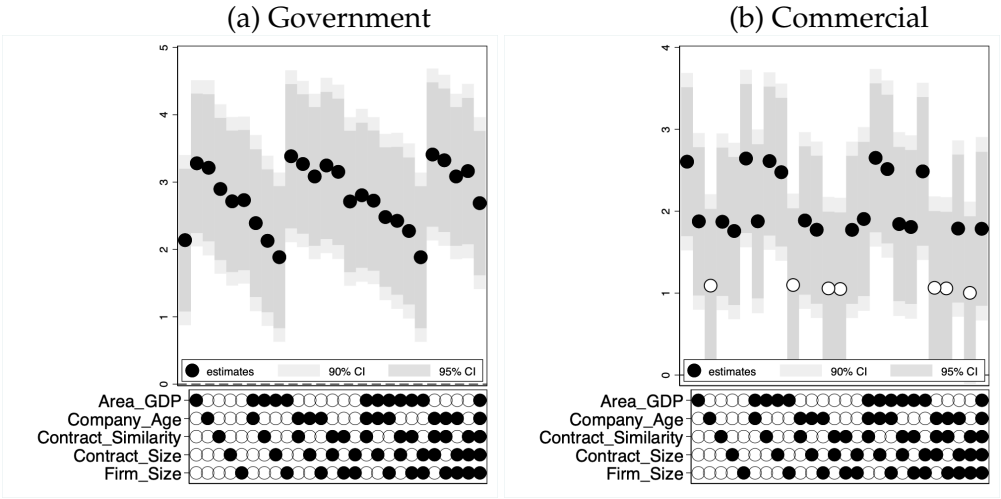
Differences in the terms and tasks under data-rich contracts One naturally wonders whether firms receiving data-rich public security contracts are engaged in similar work to firms receiving data-scarce public security contracts. We first examine whether data-rich contracts are similar in *content* to data-scarce contracts. 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 has no effect on our results (see Figure 3 and Table A.11, Panel B).

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 or 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. Two pieces of evidence suggest that learning by doing is not driving our main results. First, while learning by doing may certainly 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 was positively correlated with data-richness. This 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 data-rich public security contract is very similar to



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



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

Figure 3: Panel A: baseline results presented in Table A.5, adding various controls. Software development intended for government (left, (a)) and commercial (right, (b)) relative to the time of receiving initial procurement contract, controlling for firm and time period fixed effects. Coefficient on the interaction for public security contracts 3 years after contract receipt is presented. Panel B: baseline results presented in Appendix Table A.6, adding various controls. Software development intended for government (left, (a)) and commercial (right, (b)) relative to the time of receiving initial procurement contract, controlling for firm and time period fixed effects. Coefficient on the interaction for high surveillance capacity 3 years after contract receipt is presented. Hollow (white) dots indicate coefficients that are not significant at the 5% level.

the software produced after the receipt of a data-scarce one, suggesting that the underlying production processes should be similar as well.

Second, the possibility of learning by doing 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 baseline specification allows us to (imperfectly) account for this. In addition, we estimate our baseline specification, but now including pre-contract government software production, or 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). If learning by doing was the main driver our findings, then controlling for pre-contract software production in these sub-categories should substantially decrease the estimated impact of receiving a data-rich contract. We instead find that controlling for pre-contract sub-category software production only slightly reduces the estimated effect of a data-rich contract, which remains quantitatively large and statistically significant (see Appendix Table A.12).

Government contracts as sources of capital Another important consideration is that contracts may affect firms' software production through mechanisms other than the provision of data. One such mechanism is 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.11, Panel C). 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 $\sum_T T_{it} \times capitalization_i$). Again, adding this set of controls has no impact on our results (see Figure 3 and Appendix Table A.11, Panel D).

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.13, Panel B, 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.

As an alternative approach, we can examine the effects of firms' own *second* public security contracts: again, signaling effects should be much smaller for these contracts, but they should still provide access to valuable data. When we estimate our baseline specification, examining software production around the time of receiving a second contract, we continue to find significant effects of receiving a data-rich contract on government and commercial software production (see Appendix Table A.13, Panel C).

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 $\sum_T T_{it} \times market_i$). We add these interactions to our baseline specification and find that they do not affect our results (see Figure 3 and Appendix Table A.11, Panel E). A second possibility is that contracts with two very specific high-surveillance prefectures may disproportionately affect our results: Beijing and Shanghai. Contracts with these extremely powerful local governments may offer a range of political and economic opportunities that go far 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.14, Panel B). 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.14, Panel C).

Our empirical results thus paint a very 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 Macroeconomic implications of government data provision

In our empirical analysis of Section 5, we have observed some of the microeconomic consequences of government data provision to the private sector. The evidence points to the possibility that 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. Nor does it imply that increasing government data provision to private firms will increase welfare. There are two main reasons why the microeconomic and macroeconomic implications may diverge. The first is that government data provision may crowd-out resources used in other innovating firms, as well as resources used for consumption. The second is that, in general equilibrium, relative prices may change as well, affecting innovation and welfare.

Thus, in this section, we examine the role of the state in shaping macroeconomic outcomes with these considerations taken into account. We start by building a directed technical change model (Acemoglu, 2002) with data as input, where we incorporate the economies of scope implied by our evidence. Our goal is to analyze how growth and the direction of innovation in data-intensive economies is shaped by the two features of data we highlighted: states being key collectors and repositories of data, and government data generating economies of scope in innovation. We present the environment and characterize a balanced-growth path equilibrium in Section 6.1; we then discuss both positive and normative macroeconomic implications of government data provision in Section 6.2.

6.1 A directed technical change model with economies of scope from data

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 intermediate inputs into the production of a government good. The state purchases this government good; to be concrete, and to link the model to our empirical setting, we refer to the government good as "surveillance."[36]

As in Section 3, we assume that government data can be shared across uses within the

[36]The government good could also be thought of more generally as a set of state services involving sensitive and strategic data that are typically possessed by the state, including mapping services and the collection of sensitive health information, among others.

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.[37] Furthermore, while both types of data are excludable, we assume that only private data can be purchased in the market. In contrast, as in Section 3, government data can only be accessed by obtaining a contract for producing government software varieties for the state.

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

[37] This corresponds, for instance, to information collected from consumers when performing online transactions. It is worth noting that, by giving the rights to selling private data to a representative firm owned by the representative household, we are ignoring a number of interesting issues regarding how to allocate property rights between firms and consumers whose transactions generate data as a by-product.

duce 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 static profits given sectoral 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.[38] Dropping the j index for notational convenience, government software production uses government data d_g and intermediate goods x_g to produce a variety of quality q_g. Commercial software production uses both government and private data, d_g and d_p, as well as intermediates x_c to produce a variety of quality q_c.

Specifically, we assume that the firms may produce government and commercial software using the following technologies (which are a special case of those in introduced 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.[39] 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 consumption good (whose price is normalized to 1) and that all product varieties never depreciate.[40]

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

[38] 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.

[39] 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.

[40] 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.

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

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

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

We consider a balanced growth path with constant interest rate r where there is 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 variety innovators. Given the assumptions above and setting $\mu_g = \mu_c = 1$, a balanced growth path equilibrium where both types of software producing firms are present requires:

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

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

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

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

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

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

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

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

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 government taxes.

Data supply and the state We assume that the state purchases government software at price p_g in order produce a government good (surveillance) with linear technology $G = Y_g$. Moreover, it sets lump sum taxes T on households such that budget balance holds at each point in 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.[41] Then, given a measure N_g of government software innovators and a dataset available to them $\bar{d}_g$, we have that:

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

[41] 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 assume that government data is rival *across* firms: this can correspond to a local government collecting its own surveillance data and contracting with a specific firm to analyze it, thus restricting its use to that firm. Allowing government data to be non-rival across firms would magnify the importance of government data; by assuming rival government data, we are able to 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).

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. Then,

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.[42] Suppose it can mine $\kappa_p Y$ units of data out of Y, then the supply of private data is:[43]

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

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 private 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.[44]

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.*

[42] In reality, private data production (and government data production) would involve the use of other intermediate inputs. We have explored this and found that it does not meaningfully change our analysis.

[43] Note that this firm will be making positive profits in equilibrium, which are then redistributed to the households. 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.

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

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.[45] Proposition 1 in Appendix A.1 lays out sufficient conditions for such a BGP to exist and be unique.

We now formally define two objects that will be of interest in the next section. 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 effects of government data provision on innovation and welfare

Using the model of data-intensive technical change, we now analyze the role of the state in shaping macroeconomic outcomes in a data-intensive economy. We 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 the private sector increase the economy's growth rate and bias the direction of private innovation towards software.

Theorem 1 (Government data provision and innovation) *Assume $\epsilon \geq 1$ and the sufficient conditions in Proposition 1 for a unique BGP equilibrium to exist hold. An increase in government data provided to firms ($\bar{d}_g$) will increase the rate of innovation (η) and bias private innovation towards data-intensive software (increase n_c).*

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

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 ($\epsilon \geq 1$), this implies that the new entry of commercial software innovators will be sufficient to bias private innovation towards data-intensive software.

Finally, we note that the consequences of an increase in $\bar{d}_g$ are identical 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. This is shown in Appendix A.3. Thus, in a BGP of our model, it is not relevant whether the state's policy is set in terms of the provision of data ($\bar{d}_g$), the production of surveillance ($p_g G/Y$), or the collection of data (D_g/Y).

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 $\rho - \eta(1-\theta) > 0$ in a BGP, 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 $\bar{d}_g$:

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

The first is the crowding-out of resources that are used for creating new varieties (i.e., innovation) instead of consumption. The second is the crowding-out of resources that are instead used as intermediates for producing government surveillance instead of consump-

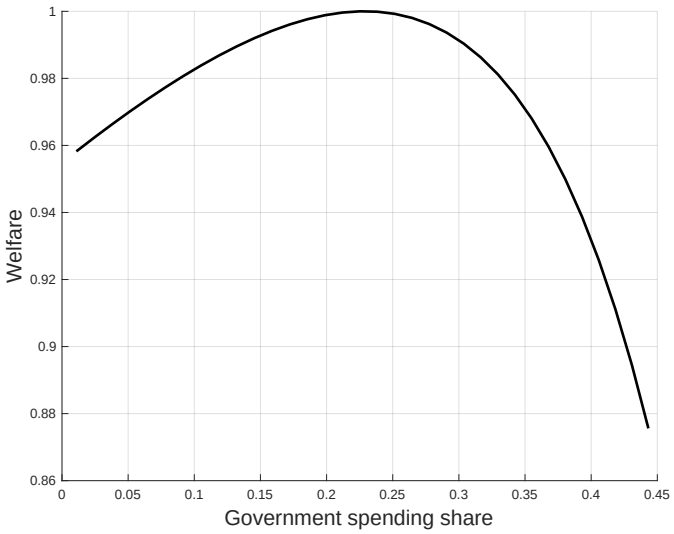


Figure 4: Government data provision and welfare.

tion.

Given this discussion, we next consider a second-best problem where the government chooses the level of government data provision to maximize household welfare.[46] 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.[47] Then, we vary the level of government data provision $\bar{d}_g$ from this benchmark parameterization, keeping all other parameters fixed.

In Figure 4, one sees that, given our parameterization of the model, the second-best government data provision results in a government spending share of 22 percent. 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 0.1, then welfare is about 2 percent lower in consumption equivalent units. The reason is that

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

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

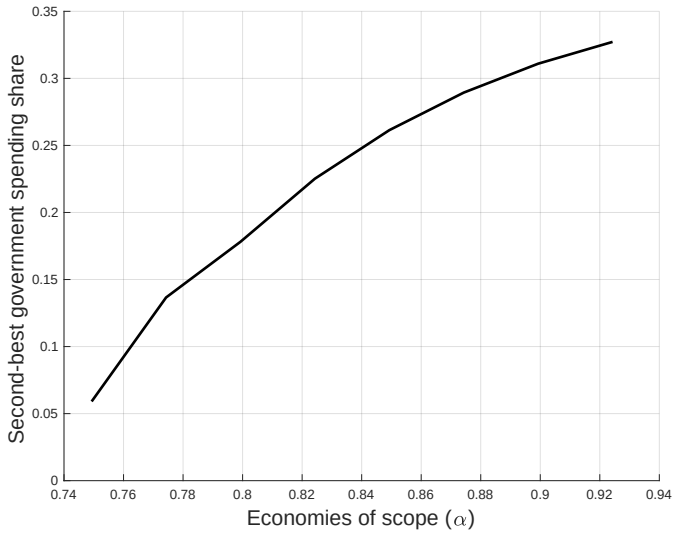


Figure 5: Economies of scope and second-best government data provision.

the growth rate η is lower, and this is not sufficiently compensated by a reduction in the crowding-out of resources for innovation and consumption. The opposite is true when government data provision is too generous: in that case, there is significant crowding out of innovation and consumption that outweighs the increase in η.

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 because the increase in the economy's growth rate is not sufficient to compensate for the crowding out of resources? To answer this, Figure 5 shows how the welfare maximizing government spending share ($\frac{p_g G}{Y+p_g G}$) changes as economies of scope become stronger (as governed by α). We find that when α is below 0.75 then it is never optimal for the state to supply government data. Therefore, when economies of scope are sufficiently low, the decentralized BGP equilibrium would only feature the production of commercial software using private data alone, and no production of government software for 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 which compensate for the crowding out of resources from consumption. In fact, when economies of scope are sufficiently strong, the state would optimally choose to provide such a high level of government data that the only software producing firms that exist in a BGP are those that produce both government and commercial software.

7 The role of the state in the age of data-intensive innovation

We have previously shown that the two features of data we highlight imply an important role for the state in shaping innovation, growth, and thus welfare. In what follows, we develop three applications illustrating that because of these features: (*i*) industrial policy in the form of government data provision is justified on different grounds from traditional industrial policy; (*ii*) surveillance states' interest in monitoring and controlling their citizens is aligned with promoting data-intensive innovation but may reduce citizen welfare; and, (*iii*) regulation limiting government data collection due to privacy concerns will reduce innovation but may increase welfare.

7.1 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.[48]

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. The justification is even stronger when government services that produce data as a by-product (like surveillance in our model) are also directly valued by either the state or households. Moreover, this process need not be limited to the Chinese context.[49]

Our model also suggests sources of variation in second-best policy choices that are particularly salient in data-intensive economies. In addition to variation in economies of scope (e.g., due to the sectoral composition of an economy), the availability of good substitutes for government data and the productivity of data production can all determine whether government provision of data to firms should be larger or smaller (or even non-existent). Differences in production technologies can thus have large effects on second-best

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

[49] For example, in the US, *Wired* writes that "the Pentagon believes AI has matured enough to become a central plank of America's national security … The plan depends on the Pentagon working closely with the tech industry to source the algorithms and cloud computing power needed to run AI projects. Federal contracting records indicate that Google, Oracle, IBM, and SAP have signaled interest in working on future Defense Department AI projects." See `https://www.wired.com/story/pentagon-doubles-down-ai-wants-help-big-tech/`, last accessed March 25, 2020.

policies. Figure 5 has already shown the effects of varying the technological parameter α governing economies of scope. As another example, we find that, when we double the elasticity of substitution γ between private and government data compared to our benchmark calibration (i.e., 6.8 versus 3.4), then the second-best government spending share falls to 0.13 from the benchmark 0.22.

Finally, we note that, even without intending to do so, different state policies may also have important industrial policy components. This echoes the argument of Rodrik (2007) that all policies, whether intended or not, can be considered as industrial policies. Our second and third applications present two such instances: the state's demand for surveillance to monitor the population, or regulations on government data collection or data sharing due privacy concerns will have important industrial policy components despite being pursued with very different objectives.

7.2 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, "A sophisticated new set of technological tools ... that will allow strongmen and police states to bolster their internal grip, undermine basic rights and spread illiberal practices beyond their own borders."[50] 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.[51]

Our model and empirical results suggest a potential alignment between surveillance states and data-intensive innovation. Greater purchases of the surveillance good G not only increase the state's political control, but also produce the government data that fuels data-intensive innovation.

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

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

[50]The *Wall Street Journal* article is: "The Autocrat's New Tool Kit," by Richard Fontaine and Kara Frederick, March 15, 2019. Available online at: `https://on.wsj.com/2H1sIgu`, last accessed August 7, 2019.

[51]Source: IHS Markit Technology Report, described by the *Wall Street Journal*, `https://on.wsj.com/2UOuuIJ`, last accessed on June 16, 2020. More generally, many democratic states have also recently been expanding their surveillance apparatus, such as the United States since the enactment of the USA PATRIOT Act of 2001.

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.[52] 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.

To illustrate the implications of variation in δ, we 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 percent lower in the US than China in 2018.[53] 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 government surveillance ($\frac{p_g G}{Y+p_g G}$) by 40 percent in our model? We find that the annual growth rate (η) decreases from the benchmark 6 percent to 4.8 percent and that the bias of innovation (n_c) decreases from the benchmark 1 to 0.82.

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 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 the *state's* optimal data-provision to firms could be very different from *citizens'* optimal data provision — states and citizens may have different values of δ. The choice of large quantities of surveillance G may increase rates of innovation and growth, but significantly reduce welfare through violations of civil liberties.

7.3 States' choice of privacy protection

States not only collect and hold data, but also regulate the collection and exchange thereof. This regulation will — especially in democracies — often reflect citizen norms regarding privacy; personal data in particular have an element of "repugnance" (Roth, 2007), with many individuals expressing discomfort when private data — especially government data — are collected and commoditized.

[52] Even selfish autocrats may value $u(C)$ when the probability of sustaining political power increases with household utility.

[53] US spending was 0.8 percent of GDP vis a vis 1.32 percent of GDP for China. Sources: "USA Spending.gov", `https://www.usaspending.gov/#/explorer/budget_function`, last accessed on April 21, 2020 and "Keynote Speech by H.E. Liu Xiaoming, Chinese Ambassador to the UK, at the Meeting with the Defence Correspondents' Association", `https://www.fmprc.gov.cn/mfa_eng/wjb_663304/zwjg_665342/zwbd_665378/t1695061.shtml`, last accessed on April 21, 2020.

While there is limited systematic evidence on citizens' privacy concerns globally, it appears that there exist large differences in consumers' willingness to pay to protect their data across the US, UK, Germany, China, and India.[54] On one end of the spectrum, Chinese consumers show low levels of privacy concerns. On the other end of the spectrum, Germans exhibit substantial privacy concerns regarding a broad range of data, reflected in, for example, the EU's strict General Data Protection Regulation.[55]

We focus here on restrictions on the state's collection and sharing of data, arising from norms of privacy. Our model suggests that the expression of norms regarding data's repugnance (ultimately reflected in regulation) can significantly affect the rate and direction of technical change. Consider an extension of our model where the flow utility is:

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

A positive ϕ captures households' repugnance towards government data production and data sharing. If households can enforce their preferences through regulations limiting data collection, production, or sharing, then a benevolent state would produce a lower level of aggregate government data in a BGP (i.e., a lower D_g/Y).[56] As a result, if such preferences vary across states, then the rate and direction of innovation will vary as well. Note that in our model, lower growth rates and less data-intensive innovation would be welfare *enhancing* for citizens who value privacy.

To illustrate the implications of variation in ϕ, 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 percent 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.[57] We find that the annual growth rate (η) decreases from the benchmark 6 percent to 4.5 percent and that the bias of innovation (n_c) decreases from the benchmark 1 to 0.78.

[54]Source: "Customer Data: Designing for Transparency and Trust", `https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust`, last accessed on February 5, 2020.

[55]Related to our previous application, a tolerance of data collection might be an explicit aim of autocrats: such tolerance would align citizen attitudes with the monitoring objective of the state (as well as with the interests of innovators relying on data). China's development of a "social score" that provides individuals with incentives to be monitored is such an expression of the Chinese Communist Party's desired alignment of norms and institutions (see, e.g., Besley and Persson, 2019, for a discussion of the interrelationship between culture and political institutions).

[56]One example of such a regulation is the General Data Protection Regulation in the European Union. It attempts to protect citizen privacy by limiting data collection, with important consequences for innovation. See Aridor et al. (2020) for a recent examination of how GDPR affects the amount and type of data available to firms.

[57]Source: "Top 10 Countries and Cities by Number of CCTV Cameras", `https://www.aithority.com/news/top-10-countries-and-cities-by-number-of-cctv-cameras/`, last accessed on April 21, 2020.

8 Conclusion

In this paper we argue that states — through their collection, distribution, and regulation of data — could be critical in shaping innovation in the coming decades as data-intensive technologies become widespread. States may play an important role in such data-intensive economies because of two features of data as an input: (*i*) throughout history and up to the present, states have been key collectors of data, and (*ii*) data is sharable across multiple uses within firms. These two features imply that economies of scope may arise from government data. We provide empirical evidence of economies of scope in the context of a prototypical data-intensive sector: facial recognition AI in China. We then build a directed technical change model to show that government data provision can increase the economy's growth rate and bias the direction of private innovation, but that this policy only increases welfare when economies of scope are sufficiently strong and when states' and citizens' preferences are sufficiently aligned.

Our analysis suggests several directions for future research. First, many uncertainties remain about the implications of government data provision as a policy to promote innovation. We have provided a theoretical justification for this policy, and evidence on one of the determinants of its consequences: economies of scope. Yet, a comprehensive quantitative assessment would require further measurement. For 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, 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.[58] 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.

[58]Consider the facial recognition AI sector as an example: Appendix Figure A.16 presents the companies with the top facial recognition algorithms, as ranked by the Face Recognition Vendor Test in 2018, organized by the National Institute of Standards and Technology, an agency of the US Department of Commerce. As one can see, Chinese firms occupy all of the top 5 positions, with Yitu, SenseTime, and MegVii, the 3 largest firms in China, ranking 1st, 3rd, and 8th, respectively (see `https://bit.ly/2OKL5Ze` for more details, last accessed on August 8, 2019). Lee (2019) documents China's leadership in several sectors of the AI industry.

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.

_ , "Equilibrium Bias of Technology," *Econometrica*, September 2007, *75* (5), 1371–1409.

_ **and J Linn**, "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry," *The Quarterly Journal of Economics*, August 2004, *119* (3), 1049–1090.

_ **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.

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

_ **, Philippe Aghion, and Fabrizio Zilibotti**, "Distance to Frontier, Selection, and Economic Growth," *Journal of the European Economic Association*, March 2006, *4* (1), 37–74.

_ **,** _ **, 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.

Aridor, Guy, Yeon-Koo Che, William Nelson, and Tobias Salz, "The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR," *Working Paper*, January 2020, pp. 1–67.

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.

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

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

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*, August 2019, *33* (3), 163–184.

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

Bronzini, Raffaello and Eleonora Iachini, "Are Incentives for R&D Effective? Evidence from a Regression Discontinuity Approach," *American Economic Journal: Economic Policy*, November 2014, *6* (4), 100–134.

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

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

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

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

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

Gates, Kelly A, *Our Biometric Future* Facial Recognition Technology and the Culture of Surveillance, NYU Press, January 2011.

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

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

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

Habakkuk, H J, *American and British Technology in the Nineteenth Century* The Search for Labour Saving Inventions, Cambridge University Press, 1962.

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

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

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

Hicks, John, *The Theory of Wages*, London: Springer, June 1932.

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

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

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

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

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

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

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

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

Lee, Kai-Fu, *AI Superpowers* China, Silicon Valley, and the New World Order, Mariner Books, 2019.

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

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

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

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

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

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

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

_ **and Alessandra Voena**, "Compulsory Licensing: Evidence from the Trading with the Enemy Act," *American Economic Review*, February 2012, *102* (1), 396–427.

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

Newell, R G, A B Jaffe, and R N Stavins, "The Induced Innovation Hypothesis and Energy-Saving Technological Change," *The Quarterly Journal of Economics*, August 1999, *114* (3), 941–975.

North, Douglass C, "Institutions," *Journal of Economic Perspectives*, February 1991, *5* (1), 97–112.

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

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

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

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

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

Rosenstein-Rodan, P N, "Notes on the Theory of the 'Big Push'," in "Economic Development for Latin America," London: Palgrave Macmillan UK, 1961, pp. 57–81.

Roth, Alvin E, "Repugnance as a Constraint on Markets," *Journal of Economic Perspectives*, July 2007, *21* (3), 37–58.

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 R W Vishny, "Corruption," *The Quarterly Journal of Economics*, August 1993, *108* (3), 599–617.

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

Simon, Brenda M and Ted Sichelman, "Data-Generating Patents," *Northwestern University Law Review*, February 2017, *111* (2), 1–62.

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

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

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

Wei, Shang-Jin, Zhuan Xie, and Xiaobo Zhang, "From "Made in China" to "Innovated in China": Necessity, Prospect, and Challenges," *NBER Working Paper*, November 2016, pp. 1–49.

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

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

ONLINE APPENDIX (NOT FOR PUBLICATION)

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

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

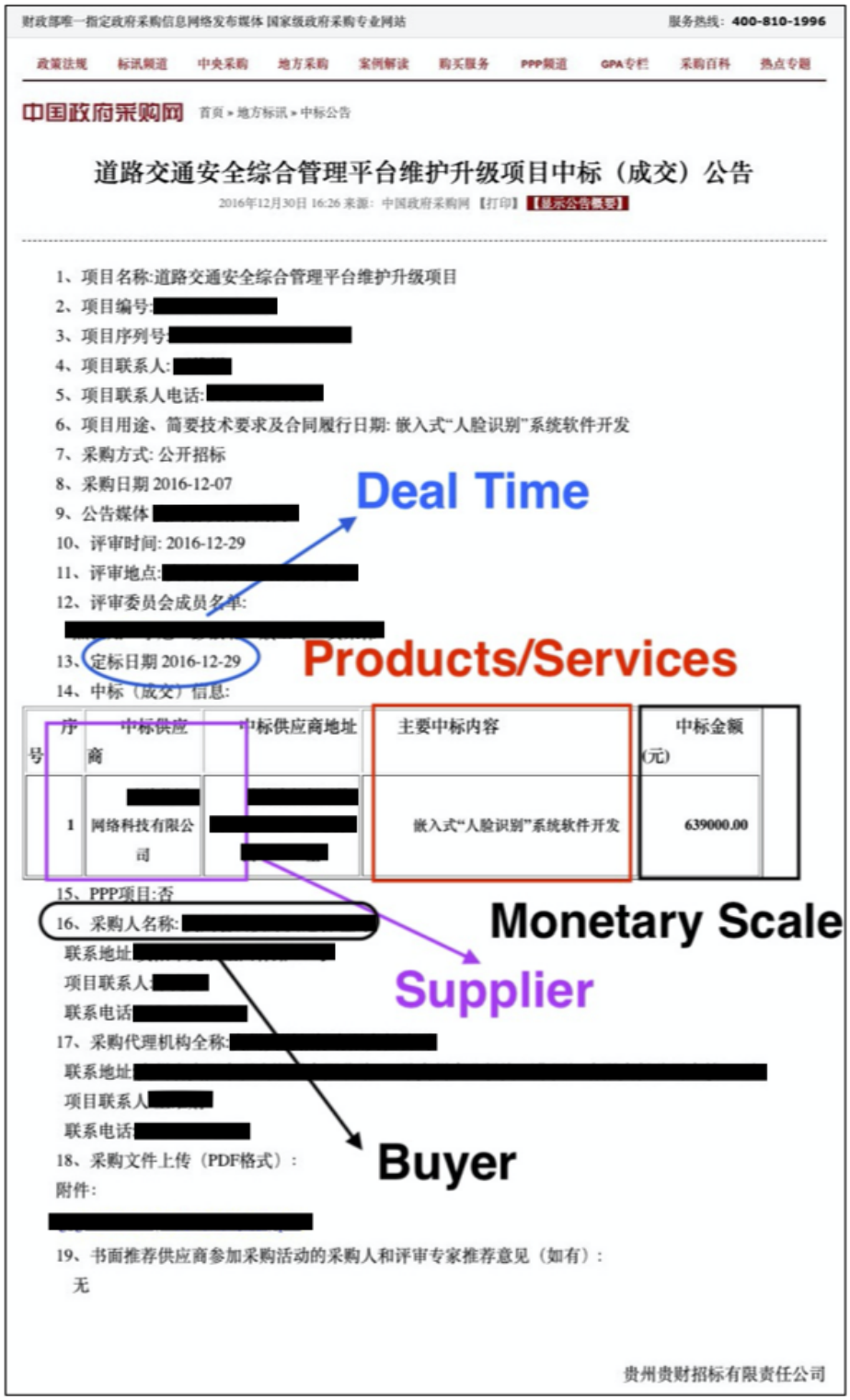
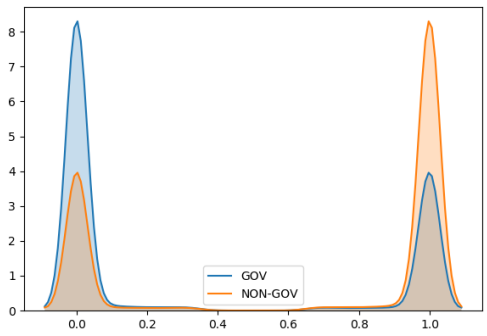
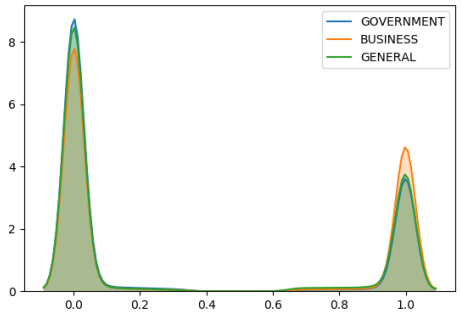


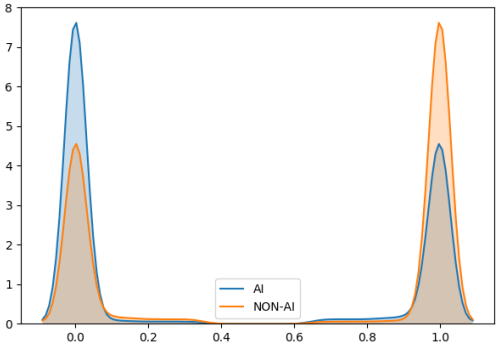
Figure A.3: Example of a procurement contract record; source: Chinese Government Procurement Database.



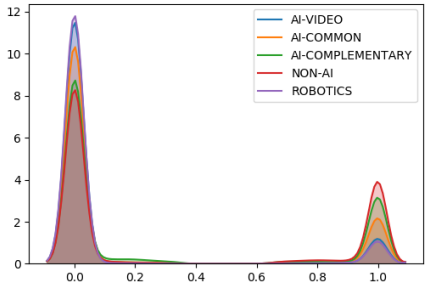
(a) Customers - Binary



(b) Customers - Non-Binary

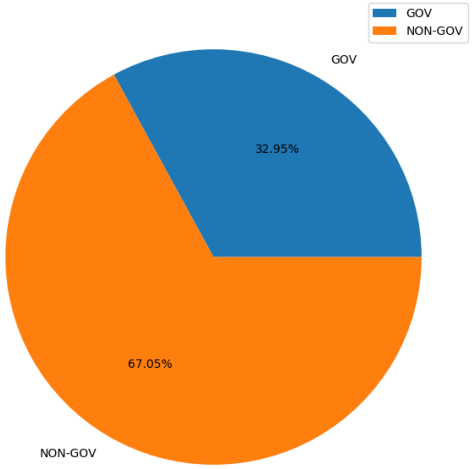


(c) Function - Binary

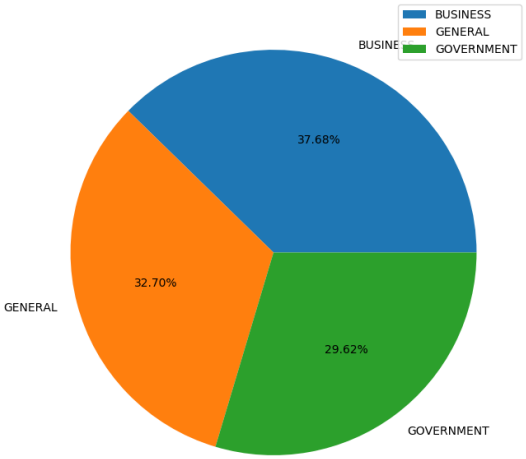


(d) Function - Non-Binary

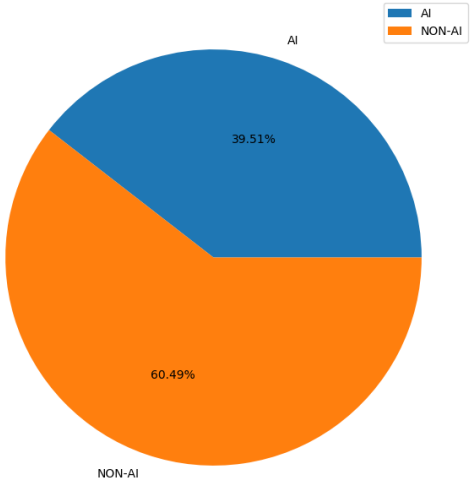
Figure A.4: Probability density plots of software categorizations based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization by customers (left is binary; right is full set of categories); Bottom panel shows categorization by function (left is binary; right is full set of categories).



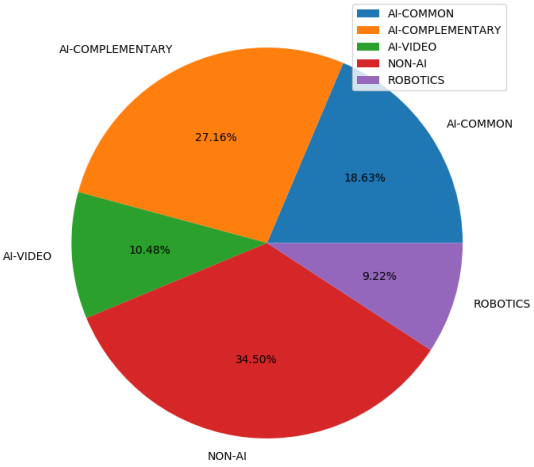
(a) Customers - Binary



(b) Customers - Non-Binary

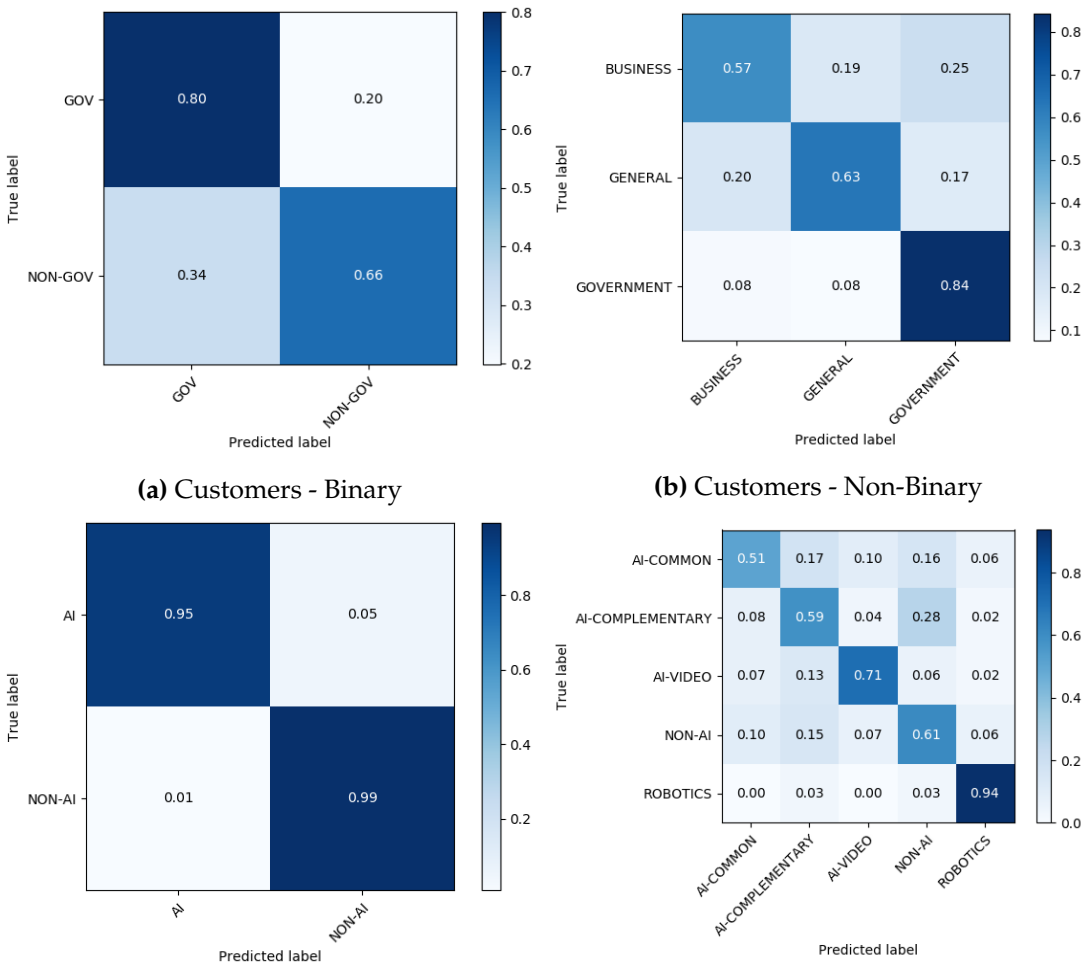


(c) Function - Binary



(d) Function - Non-Binary

Figure A.5: 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 (left is binary; right is full set of categories); bottom panel shows categorization by function (left is binary; right is full set of categories).



(a) Customers - Binary

(b) Customers - Non-Binary

(c) Functions - Binary

(d) Functions - Non-Binary

Figure A.6: 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 (left is binary; right is full set of categories); bottom panel shows categorization by function (left is binary; right is full set of categories).

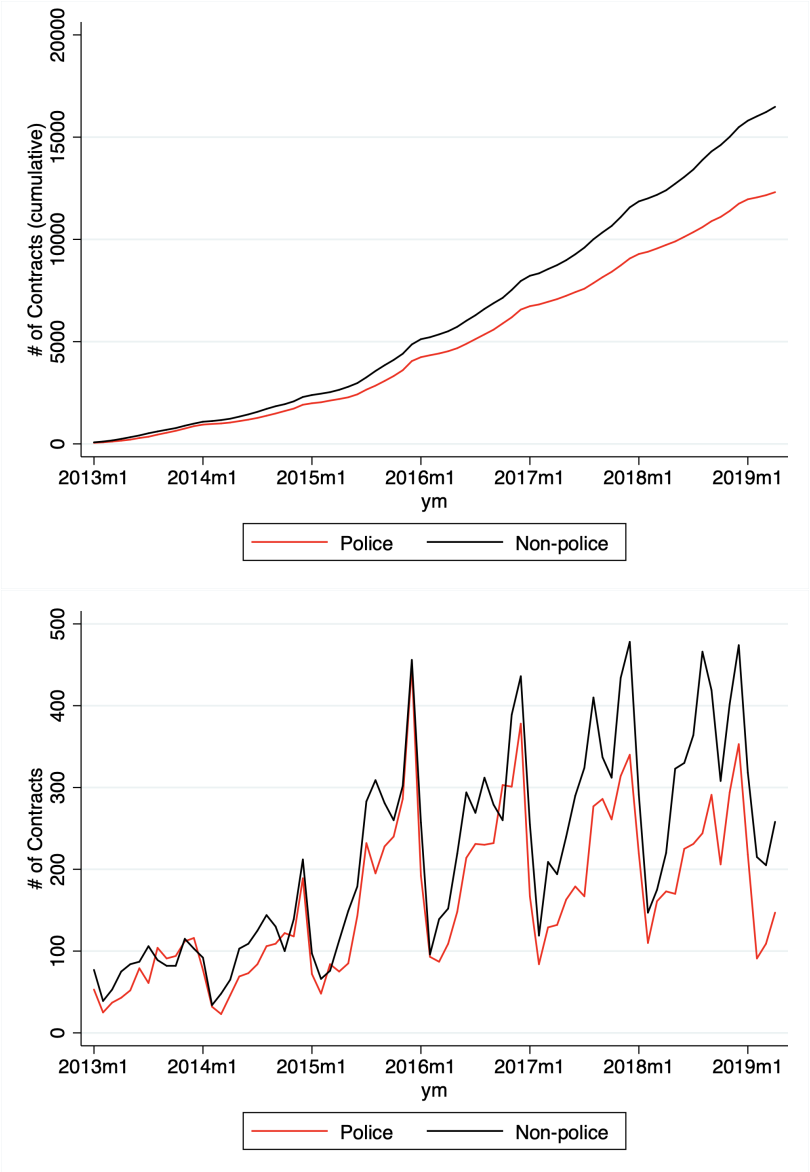


Figure A.7: Cumulative number of public security and non-public security contracts (top panel), and the flow of new contracts signed in each month (bottom panel).

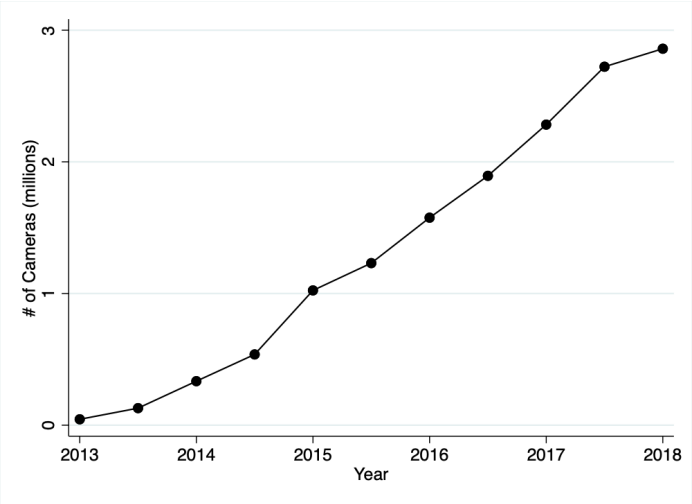
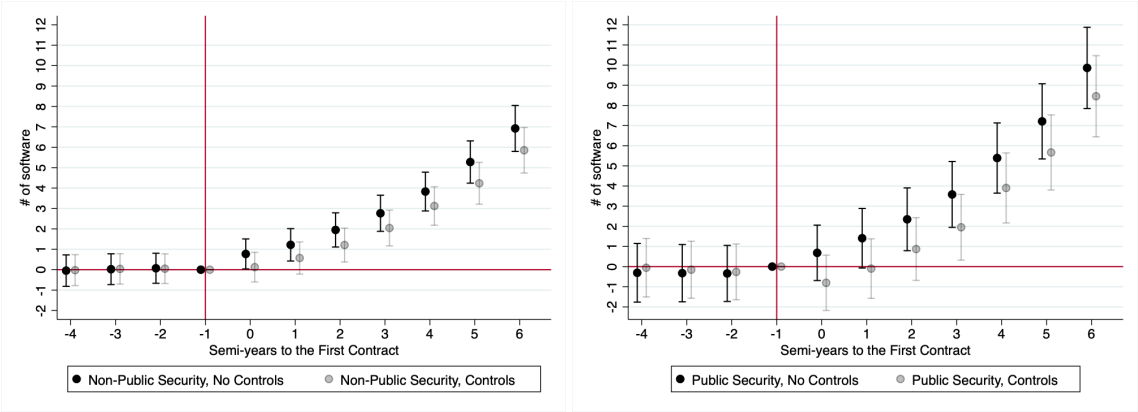
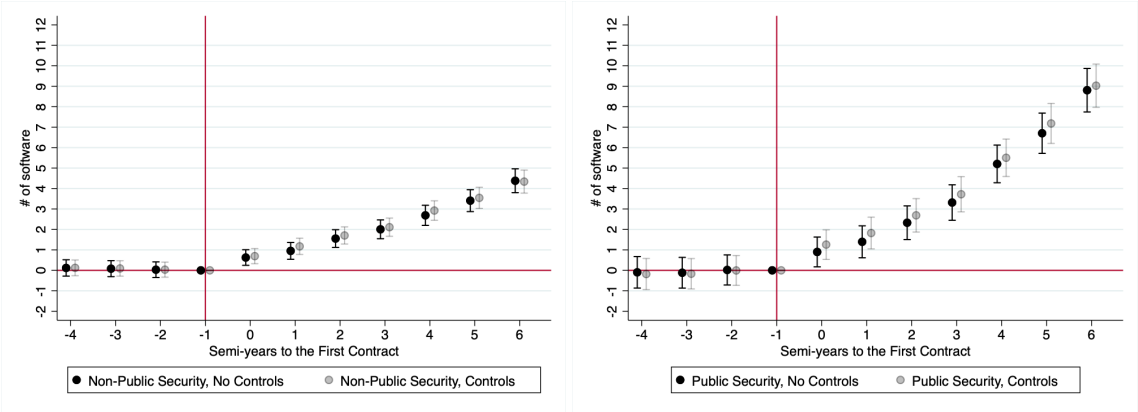


Figure A.8: Number of new public surveillance cameras in China since 2013, as measured by government procurement contracts on cameras. Source: procurement contracts database.

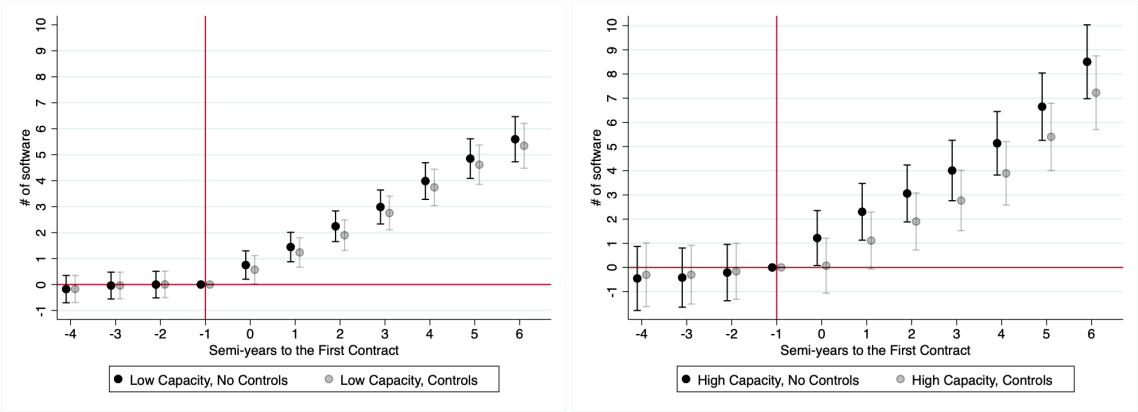


(a) Government

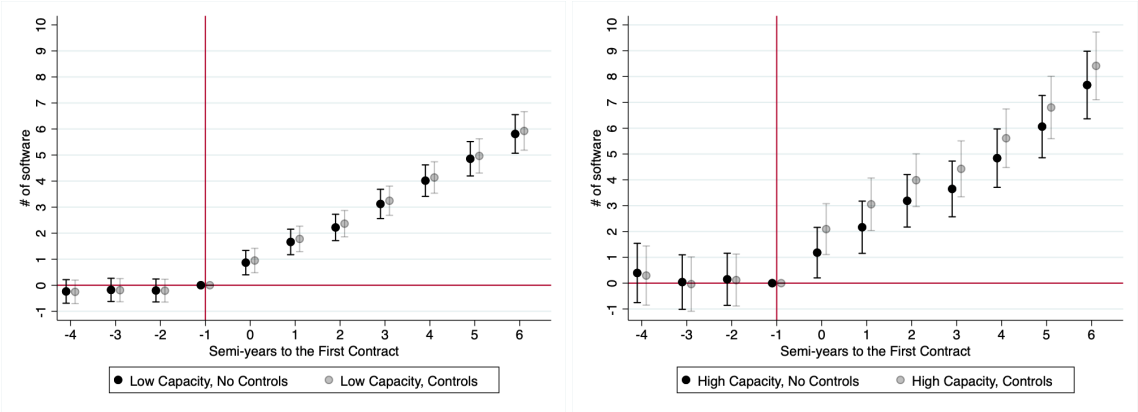


(b) Commercial

Figure A.9: 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.

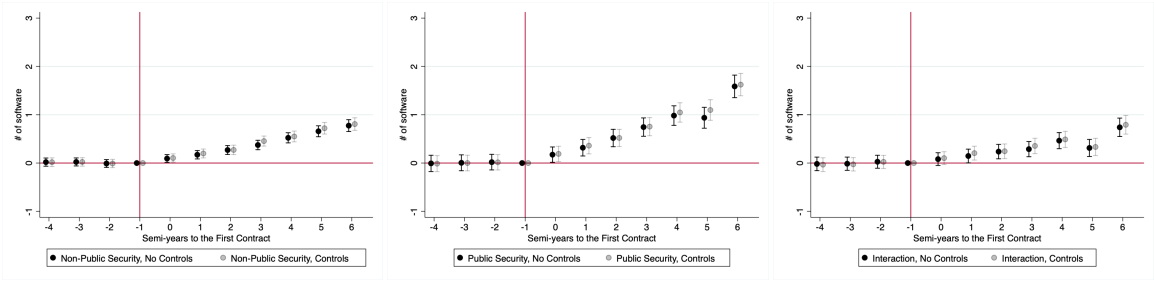


(a) Government

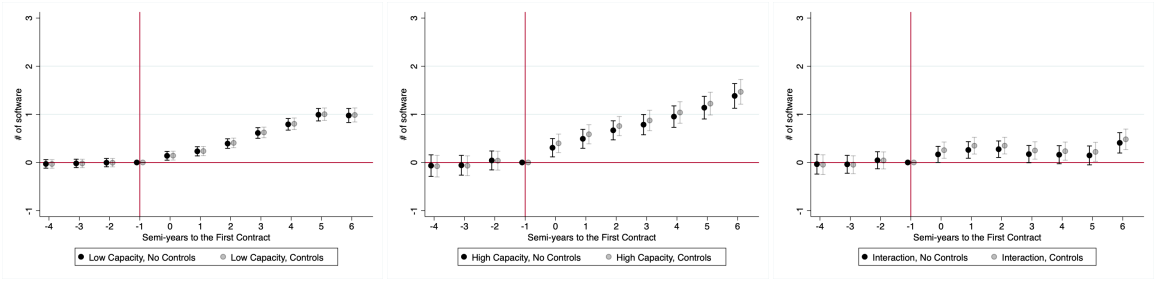


(b) Commercial

Figure A.10: Software development intended for government (Panel A) or for commercial uses (Panel B), resulting from data-rich contracts (right column) 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 A: Public security vs. non-public security contracts



Panel B: High capacity vs. low capacity security contracts

Figure A.11: 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 interaction (right column). Panel B: Results for public security contracts that are data-scarce (left column), data-rich (middle column), and the interaction (right column). All figures control for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.

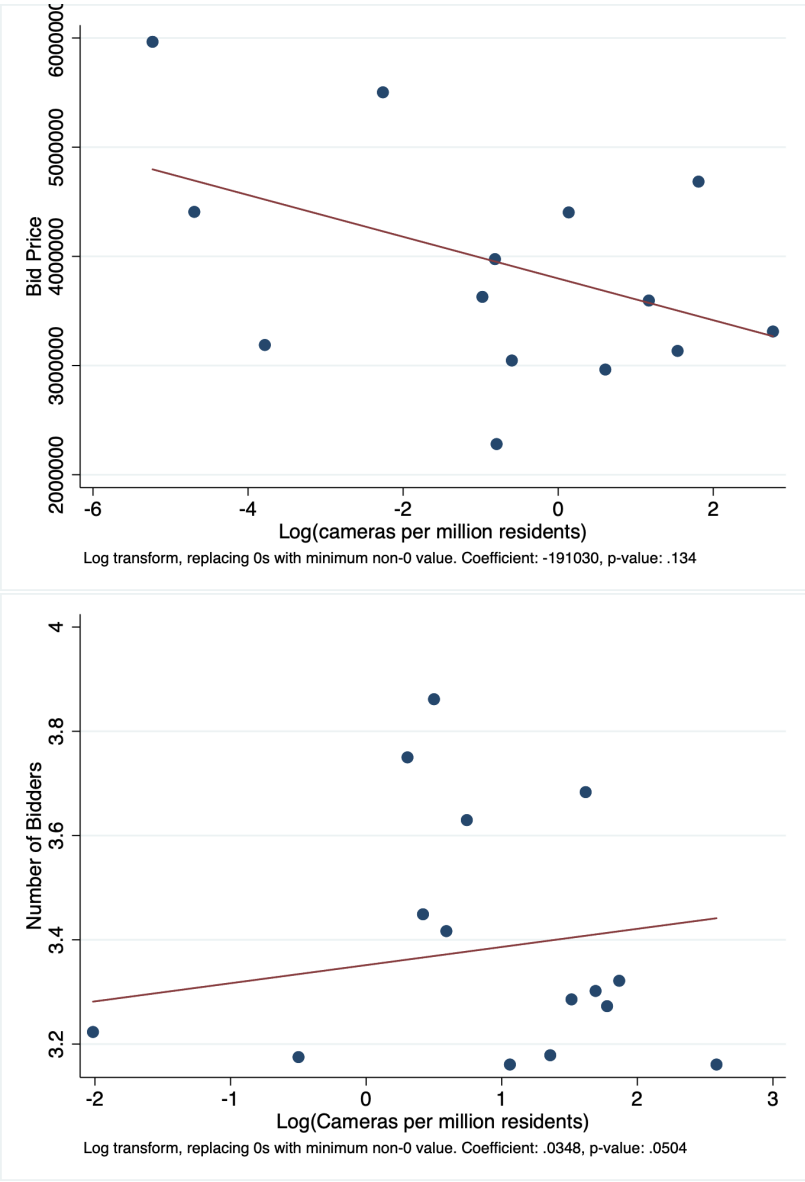
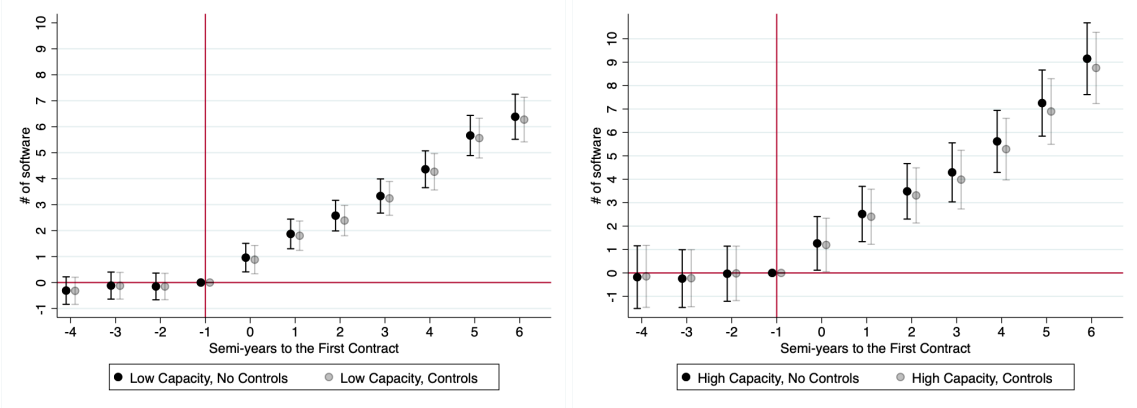
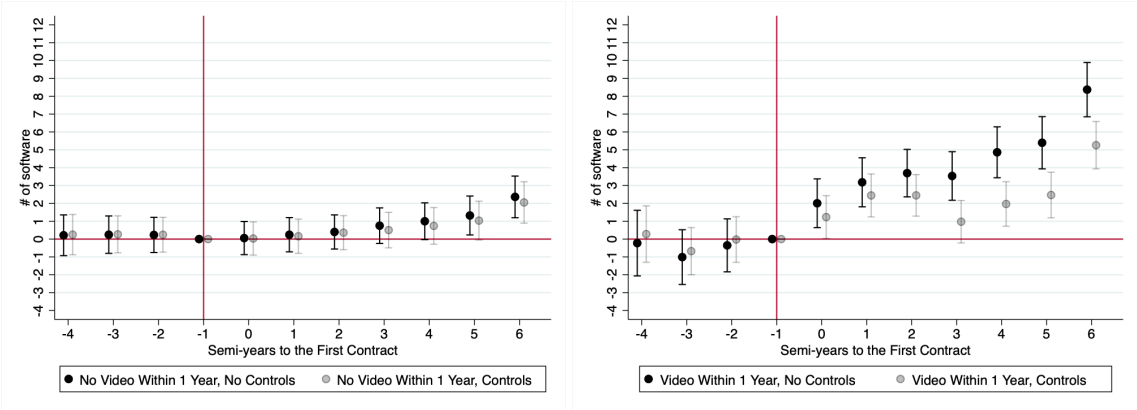


Figure A.12: Binned scatterplots of size of bid versus prefecture surveillance capacity, residualizing company fixed effects (top); and of number of bidders versus prefecture surveillance capacity (bottom).



Panel A: Data-Complementary, split by surveillance capacity



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

Figure A.13: Panel A: Data-complementary software production resulting from data-rich contracts (right column) 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.

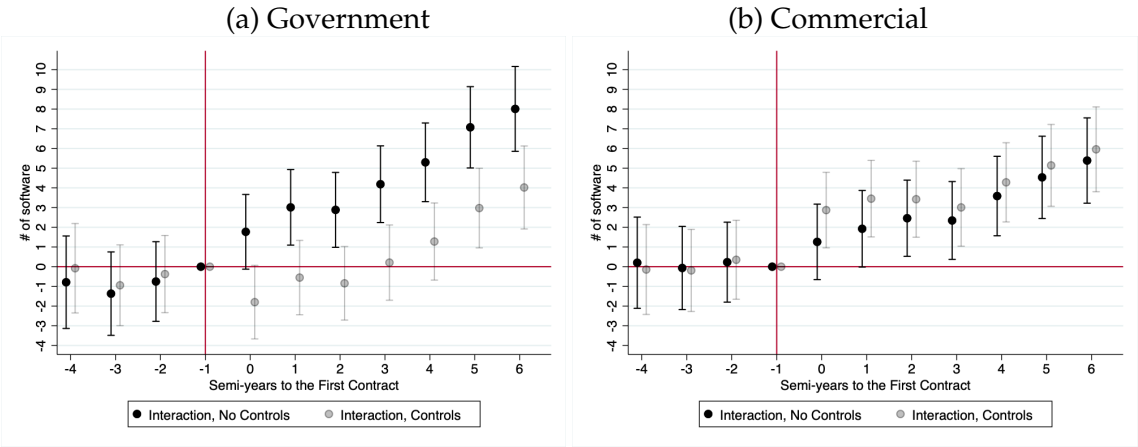


Figure A.14: Differential software development intended for government (left column) or for commercial uses (right column), resulting from public security contracts that led to data-complementary software production within 1 year, relative to public security contracts that did not, controlling for firm and time period fixed effects. Translucent lines/markers additionally interact pre-contract firm characteristics with a full set of time-period fixed effects.

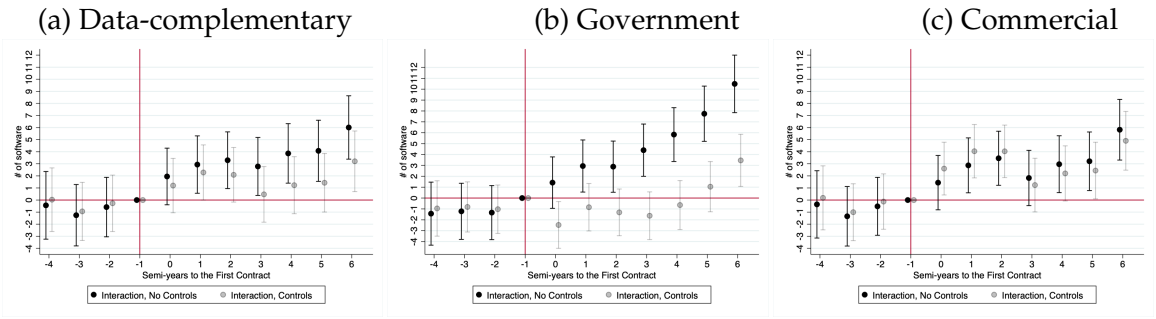


Figure A.15: Differential data-complementary software (left column) and differential AI software development intended for government (middle column) or for commercial uses (right column), resulting from public security contracts that led to video facial recognition AI software within 1 year, relative 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.

#	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 A.16: Face Recognition Vendor Test (FRVT) ranking of top facial recogntion algorithms. Source: *National Institute of Standards and Technology (NIST)*.

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 — 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 refers to spending/revenue by the city's government.

Table A.3: Top predicted words from LSTM model — binary categorization of software

Chinese Term	English Translation	Frequency (%)
(1)	(2)	(3)
Panel A: Customer is Government		
视频	Video	1.165
监控	Monitor	.925
摄像机	Video Camera	.836
识别方法	Recognition Method	.402
管理	Management	.365
云台	Tripod Head (gimbal)	.354
介质	Medium	.312
传输	Transmission	.300
人脸识别	Facial Recognition	.292
无线	Wireless	.270
人脸	Face	.254
安全	Safety	.245
公仔	Doll	.234
摄像	Video	.227
控制器	Controller	.219
进行	Execute	.209
监测	Monitor	.208
采集	Collection	.203
切换	Switch	.202
远程	Long Distance/Remote	.193
Panel B: Function is AI		
室内	Indoors	.558
语音	Voice	.463
识别	Recognition	.451
识别方法	Recognition Methods	.446
车辆	Car	.438
控制系统	Control System	.279
人脸	Face	.244
模型	Model	.220
目标	Target/Objective	.212
数据处理	Data Processing	.203
图像处理	Image Processing	.202
介质	In Medium (Media)	.190
驱动	Drive	.175
指纹	Fingerprint	.172
推荐	Recommendation	.172
电机	Motor	.167
控制器	Controller	.167
搜索	Search	.162
运动	Motion	.162
人脸识别	Facial Recognition	.160

Table A.4: Top predicted words from LSTM model — non-binary categorization of software

Panel A: Customer type

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

Panel B: Function type

Non-AI			AI-Video			Robotics		
Chinese	English	Freq. (%)	Chinese	English	Freq. (%)	Chinese	English	Freq. (%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
金蝶	Kingdee	.393	人脸	Face	1.104	焊接	Welding	.308
Thundersoft	Thundersoft	.241	深度	Depth	.321	百度	Baidu	.272
支付	Pay	.214	抓拍	Snapshot	.310	华恒	Huaheng	.252
DHC	DHC	.189	商汤	SenseTime	.287	预警系统	Warning System	.239
网易	Netease	.188	考勤	Attendance	.258	电梯	Escalator	.239
PC	PC	.165	科达	Kedacom	.258	生产	Production	.202
广告	Advertisement	.158	跟踪	Track	.249	传感器	Sensor	.186
用户	User	.155	全景	Panoramic	.224	瑞斯	Rees	.182
营销	Marketing	.154	广电	Broadcastt	.209	AGV	AGV	.180
数据分析	Data Analysis	.136	目标	Target/Objective	.189	无人	Unmanned	.176
项目	Project	.126	车牌	License Plate	.189	恒润	Hirain	.176
安卓版	Android	.125	特征	Feature	.184	经纬	Lat/Long	.166
客户	Client	.120	铂亚	Platinum	.175	测试软件	Test Software	.159
助手	Assistant	.119	预警	Warning	.166	小米	Xiaomi	.159
公交	Public Transport	.118	运通	American Express	.163	新时	New Time	.156
联迪	Landi	.114	指挥	Command	.158	紫光	Tsinghua	.156
桌面	Desktop	.110	统计	Statistics	.149	机械	Mechanical	.152
KIS	KIS	.108	安居	Safety	.146	进化	Evolution	.152
分众	Focus	.107	SDK	SDK	.141	交通信号	Traffic Signal	.149
android	android	.106	布控	Deploymentt	.141	遥控	Remote	.133

AI-Common			Data-Complementary		
Chinese	English	Freq. (%)	Chinese	English	Freq. (%)
指纹	Fingerprint	.342	存储	Storage	.206
训练	Training	.203	可视化	Visualization	.167
管家	Housekeeper	.201	一体化	Integration	.164
文本	Text	.151	分布式	Distributed	.162
高速公路	Highway	.150	仿真	Simulation	.157
虹膜	Iris	.147	医学影像	Medical Imaging	.148
汽车	Car	.143	通用	General	.144
海尔	Haier	.137	集成	Integrated	.141
WPS	WPS	.134	数据管理	Data Management	.136
翻译	Translate	.126	宇视	UTV	.136
推荐	Recommend	.124	管控	Manage	.126
图片	Image	.119	高速	High Speed	.126
测量	Test	.116	媒体	Media/Medium	.125
征信	Credit	.111	手机软件	Phone Software	.125
指纹识别	Fingerprint Recognition	.106	设计	Design	.117
作业	Operation	.106	接口	Interface	.117
微信	WeChat	.105	开发	Development	.116
评估	Assessment	.105	服务器	Server	.116
灵云	AIcloud	.102	处理软件	Processing Software	.113
活体	Living Body	.098	传输	Transmission	.111

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	-0.118	0.009	-0.113	-0.016	0.019	-0.014
	(0.195)	(0.138)	(0.158)	(0.217)	(0.149)	(0.176)
3 Semiyears Before	-0.117	0.003	-0.070	-0.056	0.012	-0.008
	(0.191)	(0.135)	(0.154)	(0.213)	(0.146)	(0.172)
2 Semiyears Before	-0.104	-0.033	-0.073	-0.083	-0.006	-0.039
	(0.187)	(0.132)	(0.151)	(0.208)	(0.143)	(0.169)
Receiving 1st Contract	0.399**	0.429***	0.464***	-0.120	0.489***	0.289*
	(0.186)	(0.132)	(0.151)	(0.208)	(0.143)	(0.169)
1 Semiyear After	0.891***	0.867***	1.030***	0.243	0.975***	0.710***
	(0.201)	(0.142)	(0.162)	(0.223)	(0.153)	(0.181)
2 Semiyears After	1.405***	1.372***	1.529***	0.845***	1.510***	1.277***
	(0.210)	(0.149)	(0.170)	(0.234)	(0.160)	(0.189)
3 Semiyears After	2.171***	1.985***	2.243***	1.555***	2.115***	2.060***
	(0.223)	(0.158)	(0.179)	(0.247)	(0.170)	(0.200)
4 Semiyears After	3.019***	2.632***	3.207***	2.403***	2.815***	3.038***
	(0.238)	(0.168)	(0.192)	(0.264)	(0.181)	(0.214)
5 Semiyears After	3.845***	2.958***	3.940***	3.495***	3.285***	3.845***
	(0.256)	(0.181)	(0.207)	(0.284)	(0.194)	(0.231)
6 Semiyears After	4.839***	3.748***	4.786***	4.607***	4.028***	4.734***
	(0.278)	(0.196)	(0.225)	(0.308)	(0.211)	(0.250)
4 Semiyears Before × Public Security	-0.143	-0.128	-0.077	-0.150	-0.127	-0.104
	(0.316)	(0.224)	(0.255)	(0.352)	(0.241)	(0.285)
3 Semiyears Before × Public Security	-0.049	-0.132	-0.071	-0.070	-0.132	-0.089
	(0.310)	(0.220)	(0.251)	(0.346)	(0.237)	(0.280)
2 Semiyears Before × Public Security	0.020	0.063	0.008	0.038	0.044	0.000
	(0.305)	(0.216)	(0.246)	(0.340)	(0.233)	(0.276)
Receiving 1st Contract × Public Security	0.413	0.468**	0.392	-0.279	0.551**	-0.015
	(0.301)	(0.213)	(0.243)	(0.335)	(0.230)	(0.272)
1 Semiyear After × Public Security	0.810**	0.769***	0.605**	0.296	0.919***	0.434
	(0.321)	(0.227)	(0.259)	(0.357)	(0.245)	(0.290)
2 Semiyears After × Public Security	1.433***	1.313***	1.158***	0.778**	1.241***	0.993***
	(0.336)	(0.238)	(0.272)	(0.374)	(0.257)	(0.304)
3 Semiyears After × Public Security	1.818***	1.638***	1.716***	1.259***	1.653***	1.347***
	(0.350)	(0.249)	(0.283)	(0.390)	(0.268)	(0.316)
4 Semiyears After × Public Security	2.406***	2.588***	2.302***	1.877***	2.575***	2.196***
	(0.370)	(0.262)	(0.299)	(0.411)	(0.282)	(0.334)
5 Semiyears Before × Public Security	2.989***	3.540***	3.421***	2.241***	3.512***	3.103***
	(0.391)	(0.276)	(0.316)	(0.435)	(0.297)	(0.352)
6 Semiyears After × Public Security	4.204***	4.458***	4.740***	3.537***	4.466***	4.465***
	(0.420)	(0.296)	(0.339)	(0.467)	(0.319)	(0.378)

Notes: Baseline specification (Columns 1–4) 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 below the coefficients. * 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	-0.171	-0.254	-0.319
	(0.268)	(0.231)	(0.270)	(0.267)	(0.230)	(0.267)
3 Semiyears Before	-0.040	-0.180	-0.118	-0.037	-0.190	-0.124
	(0.264)	(0.228)	(0.266)	(0.262)	(0.227)	(0.262)
2 Semiyears Before	-0.002	-0.202	-0.151	0.006	-0.209	-0.153
	(0.261)	(0.225)	(0.262)	(0.260)	(0.224)	(0.259)
Receiving 1st Contract	0.750***	0.868***	0.959***	0.569**	0.949***	0.881***
	(0.279)	(0.239)	(0.280)	(0.277)	(0.239)	(0.277)
1 Semiyear After	1.443***	1.663***	1.871***	1.236***	1.779***	1.803***
	(0.289)	(0.250)	(0.291)	(0.288)	(0.250)	(0.288)
2 Semiyears After	2.243***	2.219***	2.576***	1.897***	2.365***	2.387***
	(0.301)	(0.258)	(0.301)	(0.300)	(0.258)	(0.298)
3 Semiyears After	2.986***	3.122***	3.331***	2.755***	3.245***	3.240***
	(0.334)	(0.287)	(0.336)	(0.332)	(0.286)	(0.332)
4 Semiyears After	3.984***	4.017***	4.362***	3.740***	4.138***	4.265***
	(0.360)	(0.309)	(0.362)	(0.358)	(0.308)	(0.358)
5 Semiyears After	4.849***	4.857***	5.662***	4.614***	4.967***	5.561***
	(0.389)	(0.337)	(0.395)	(0.386)	(0.336)	(0.390)
6 Semiyears After	5.595***	5.811***	6.383***	5.342***	5.926***	6.274***
	(0.444)	(0.378)	(0.443)	(0.441)	(0.377)	(0.438)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130	-0.133	0.549	0.172
	(0.620)	(0.539)	(0.627)	(0.616)	(0.537)	(0.620)
3 Semiyears Before × High Capacity	-0.379	0.222	-0.124	-0.263	0.154	-0.101
	(0.565)	(0.488)	(0.570)	(0.561)	(0.486)	(0.563)
2 Semiyears Before × High Capacity	-0.209	0.351	0.118	-0.165	0.330	0.135
	(0.535)	(0.463)	(0.540)	(0.531)	(0.462)	(0.534)
Receiving 1st Contract × High Capacity	0.465	0.314	0.303	-0.497	1.144***	0.314
	(0.508)	(0.438)	(0.512)	(0.511)	(0.442)	(0.512)
1 Semiyear After × High Capacity	0.858	0.502	0.645	-0.126	1.276***	0.597
	(0.524)	(0.451)	(0.528)	(0.527)	(0.455)	(0.527)
2 Semiyears After × High Capacity	0.817	0.969**	0.909*	0.003	1.622***	0.924*
	(0.520)	(0.449)	(0.524)	(0.521)	(0.452)	(0.522)
3 Semiyears After × High Capacity	1.023*	0.526	0.963*	0.013	1.181**	0.745
	(0.544)	(0.470)	(0.549)	(0.545)	(0.472)	(0.547)
4 Semiyears After × High Capacity	1.151**	0.823*	1.256**	0.153	1.473***	1.022*
	(0.565)	(0.487)	(0.570)	(0.566)	(0.489)	(0.568)
5 Semiyears Before × High Capacity	1.800***	1.205**	1.592***	0.786	1.835***	1.334**
	(0.594)	(0.515)	(0.602)	(0.595)	(0.517)	(0.599)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***	1.884***	2.487***	2.481***
	(0.642)	(0.550)	(0.644)	(0.641)	(0.552)	(0.640)

Notes: All regressions estimated on the sample of firms with first contracts with a public security agency. Baseline specification (Columns 1–4) 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 below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.7: Robustness — ambiguous public security agencies

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: Dropping ambiguous companies			
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: Panel A replicates baseline specification in Table A.6, Columns 1-3; specification includes full set of time indicators and interactions with public security contracts; only selected coefficient estimates are presented. Ambiguous public security contracts are ones that contain 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. Panel B drops companies whose first contract is an ambiguous contract. Standard errors clustered at mother firm level are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.8: Robustness — LSTM categorization model configuration

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline (timestep 20, embeddings 32, nodes 32)			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: 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 C: 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 D: 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)

Notes: Specification includes 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 below the coefficients. Panel A replicates baseline specification in Table A.6, Columns 1-3 using the default LSTM specification with a timestep (phrase length) of 20, embedding size (number of dimensions in a vector to represent a phrase) of 32, and 32 nodes in the model. Panel B presents results for the same model in Panel A trained with a timestep of 10 instead; Panel C presents results for the same model in Panel A trained with an embedding size of 16 instead; Panel D presents results for the same model in Panel A trained with 16 nodes instead. The full set of combinations of results with varied model parameters do not look qualitatively different. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.9: Robustness — LSTM categorization model threshold

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline (threshold 50%)			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: 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 After × High Capacity	2.811***	2.349***	2.765***
	(0.562)	(0.740)	(0.609)
Panel C: 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)

Notes: Specification includes 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 below the coefficients. Panel A replicates baseline specification in Table A.6, Columns 1-3 using 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 and C replicate the exercise setting the threshold to be higher, at 60% and 70% respectively. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.10: Robustness — time frame

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: 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: 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: Quarter as unit of analysis			
8 Quarters Before	-0.044	-0.130	-0.062
	(0.256)	(0.214)	(0.250)
12 Quarters After	5.585***	5.436***	5.762***
	(0.416)	(0.347)	(0.403)
8 Quarters Before × High Capacity	-0.515	0.280	-0.323
	(0.570)	(0.485)	(0.560)
12 Quarters After × High Capacity	2.100***	1.516***	2.331***
	(0.585)	(0.492)	(0.571)
Subsidiary Firm FE	Y	Y	Y

Notes: Specification includes 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 below the coefficients. Panel A replicates baseline specification in Table A.6, Columns 1-3; Panel B 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 extends the time frame to 5 semi-years before and 8 semi-years after the initial contracts; Panel D uses quarters of a year as the unit of analysis, instead of semi-years. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.11: Firm, contract, and locality characteristics interacting with time trends

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: Contract similarity with opposite capacity contracts			
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 C: 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 D: 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 E: First contract location 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 F: 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 G: All pre-controls interacted			
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)

Notes: Specification includes 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 below the coefficients. Panel A replicates baseline specification in Table A.6. Panels B - G interact controls with time dummies, where Panel B interacts the monetary size of the firm, Panel C interacts the size of the contract, Panel D interacts contract similarity, Panel E interacts the GDP of the first contract's location, Panel F interacts firm age, and Panel G interacts with all the above controls. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.12: Learning by doing

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: 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 C: 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 D: 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)

Notes: Specification includes 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 below the coefficients. Panel A replicates baseline specification in Table A.6, Columns 1-3. Panel B controls for the total amount of government software produced by the firm at 1 semiyear before the contract. Panel C controls for the total of amount of software indicated in the column by the firm at 1 semiyear before the contract. Panel D 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. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.13: Effects of 2nd public security contracts

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: Sample — not first 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 C: Sample — second contract within firm			
4 Semiyears Before	-1.577*	2.214***	2.015***
	(0.916)	(0.656)	(0.697)
6 Semiyears After	8.533***	7.856***	13.538***
	(1.430)	(1.025)	(1.088)
4 Semiyears Before × High Capacity	1.090	-1.943**	-1.819*
	(1.287)	(0.923)	(0.980)
6 Semiyears After × High Capacity	29.042***	2.876**	17.833***
	(1.881)	(1.349)	(1.432)

Notes: specification includes 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 below the coefficients. Panel B 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 B from 17,400 in Panel A. Panel C replicates Panel A, except uses second contracts instead of first contracts earned by the firm. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.14: Access to commercial opportunities

	Government	Commercial	Data-complementary
	(1)	(2)	(3)
Panel A: Baseline			
4 Semiyears Before	-0.177	-0.239	-0.310
	(0.268)	(0.231)	(0.270)
6 Semiyears After	5.595***	5.811***	6.383***
	(0.444)	(0.378)	(0.443)
4 Semiyears Before × High Capacity	-0.279	0.633	0.130
	(0.620)	(0.539)	(0.627)
6 Semiyears After × High Capacity	2.911***	1.861***	2.766***
	(0.642)	(0.550)	(0.644)
Panel B: Drop Beijing, 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 C: 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: Specification includes 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 below the coefficients. Panel A replicates baseline specification in Table A.6, Columns 1-3. Panel B excludes contracts from Beijing and Shanghai (the two highest capacity prefectures/provinces), and Panel C 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 both types of firms

Proposition 1 (Existence and Uniqueness) *Let $p_z(p_c)$ be the implicit function defined by the pricing equation* (4) *and $\underline{p}_c$ the smallest p_c such that $p_z(\underline{p}_c) \geq 0$. Let $p_d(p_c)$ be the implicit function defined by*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Equations (23) and (24) then follow directly from the free-entry conditions of private

[1] Note that, as for the case of government data, we assume that private data is rival across firms. This can be seen from (32). Again, we abstract from the non-rivalry of data across firms to transparently focus on the implications of the sharability of data across uses within a firm.

innovators and market clearing in the government data and goods markets. It is easy to show that $p_z(p_c)$ exist and has a negative derivative, and that $p_d(p_c)$ also exists and has a positive derivative since profit functions are increasing in its output price and decreasing in the data input price.

Then, from the market clearing conditions in the private data and goods markets, we obtain $\tilde{N}_c/N_z, N_g/N_z$ as a function of p_c alone.

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

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,

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

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

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

To conclude the proof, we need to show conditions under which p_c exists and is unique. From the free-entry conditions for government firms, together with that of private firms, we obtain one equation that implicitly defines p_c

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

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

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

The second equality follows from the implicit function $p_d(p_c)$, the second equation from the envelope theorem, and the third equation from equation (35). The last inequality follows from (34) and the fact that $\frac{\partial \Pi_z(p_z)}{\partial p_z} \frac{\partial p_z}{\partial p_c} < 0$. It is easy to see that when $\gamma \geq 1 + \beta(\chi - 1)$, the $d_p(\bar{d}_g, p_c, p_d)$ 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 lowest p_c is lower than F/λ. A sufficient condition for this to occur is equation (26), which completes the proof.

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

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

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 private goods market we have

$$\begin{aligned}
\frac{Y_c}{Y_z} &= \frac{\tilde{N}_c}{N_z} \frac{1}{q_z(p_z)^{\frac{\chi-1}{\chi}}} q_c(0, p_c, p_d)^{\frac{\chi-1}{\chi}} + \frac{N_g}{N_z} \frac{1}{q_z(p_z)^{\frac{\chi-1}{\chi}}} q_c(\bar{d}_g, p_c, p_d)^{\frac{\chi-1}{\chi}} \\
&= \frac{N_c}{N_z} \frac{q_c(0, p_c, p_d)^{\frac{\chi-1}{\chi}} + q_c(\bar{d}_g, p_c, p_d)^{\frac{\chi-1}{\chi}}}{q_z(p_z)^{\frac{\chi-1}{\chi}}} \\
&\qquad \Pi_c(0, p_c, p_d)/p_c + \Pi_c(\bar{d}_g, p_c, p_d)/p_c \frac{1+\beta(\chi-1)}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}} \\
&= \frac{N_c}{N_z} \frac{}{\Pi_z(p_z)/p_z} \\
&= \frac{p_z}{p_c} \frac{N_c}{N_z} \left(\mu_z + \frac{\Pi_c(\bar{d}_g, p_c, p_d)}{\Pi_z(p_z)} \frac{1+\beta(\chi-1)}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}} \right) \\
&= \frac{p_z}{p_c} \frac{N_c}{N_z} \left(\mu_z + 2\frac{\Pi_c(\bar{d}_g, p_c, p_d)}{\Pi_c(\bar{d}_g, p_c, p_d) + \Pi_g(\bar{d}_g, p_c, p_d) - \frac{F}{\lambda}} \frac{1+\beta(\chi-1)}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}} \right)
\end{aligned}$$

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 $\gamma > 1 + \beta(\chi-1)$. Then, we need to show that $\frac{\Pi_c(\bar{d}_g, p_c, p_d)}{\Pi_c(\bar{d}_g, p_c, p_d) + \Pi_g(\bar{d}_g, p_c, p_d) - \frac{F}{\lambda}}$ also decreases. We have that derivative is

$$\frac{\left(\Pi_g(\bar{d}_g, p_c, p_d) - \frac{F}{\lambda}\right) \frac{\partial}{\partial \bar{d}_g} \Pi_c(\bar{d}_g, p_c(\bar{d}_g), p_d(\bar{d}_g)) - \Pi_c(\bar{d}_g, p_c, p_d) \frac{\partial}{\partial \bar{d}_g} \Pi_g(\bar{d}_g, p_c(\bar{d}_g), p_d(\bar{d}_g))}{(\Pi_c(\bar{d}_g, p_c, p_d) + \Pi_g(\bar{d}_g, p_c, p_d) - \frac{F}{\lambda})^2}$$

We know that $\frac{\partial}{\partial \bar{d}_g}\Pi_c(0, p_c(\bar{d}_g), p_d(\bar{d}_g)) > 0$, since $\Pi_z(p_z)$ increases with $\bar{d}_g$ and, in a BGP, this is equal to $\frac{1}{\mu_z}\Pi_c(0, p_c, p_d)$. Then it has to be that $\frac{\partial}{\partial \bar{d}_g}\Pi_c(\bar{d}_g, p_c(\bar{d}_g), p_d(\bar{d}_g)) > 0$, since you also have the direct effect of $\bar{d}_g$. Then, a sufficient condition for the derivative to be negative is that in the initial equilibrium.

$$\lambda \Pi_g(\bar{d}_g, p_g) < F$$

This later holds whenever equation (26) holds.

Finally, from the goods demand equations

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

Thus, when $\epsilon \geq 1$, $\gamma > 1 + \beta(\chi - 1)$ and $\lambda \Pi_g(\bar{d}_g, p_g) < F$, we then have that

$$\frac{\partial \frac{N_c}{N_z}}{\partial \bar{d}_g} > 0$$

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, because of both the higher p_g and productivity due the increase in $\bar{d}_g$. Second, through higher revenues from private software production, due to higher productivity when government data is shared across uses.

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, the direct effect of the increase in $\bar{d}_g$ dominates the second round, general equilibrium effects of the changes in prices so that the overall change in η goes in the same direction than the one determined by such direct effect.

Note that the above construction determines p_c, p_z, p_d, r and η as implicit functions of $\bar{d}_g, p_g$ purely from the free-entry conditions of firms and the Euler equation for households.

Next, we turn to the market clearing conditions to understand the change in the direction of private innovation n_c. From the definition of n_c together with equations (29) and (30), we obtain:

$$n_c = \underbrace{\left(\frac{1-a}{a}\frac{p_z}{p_c}\right)^{\epsilon}}_{=\frac{Y_c}{Y_z}} \frac{q_z(p_z)^{\frac{\chi-1}{\chi}}}{q_c(0,p_c,p_d)^{\frac{\chi-1}{\chi}} + q_c(\bar{d}_g,p_c,p_d)^{\frac{\chi-1}{\chi}}} \tag{38}$$

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 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 relative demand is sufficiently elastic ($\epsilon \geq 1$) and the conditions for a BGP to exist and be unique are satisfied, then the demand effect dominates and n_c increases.

To conclude the intuition for the theorem, consider the market clearing condition for government data (33). When $\bar{d}_g$ is higher, more government data needs to be supplied to those firms obtaining government contracts. Yet, at the old p_g, the increase in government software production 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, in a BGP, it has to be that p_g increases as well so that $q_g(\bar{d}_g, p_g)^{1-\frac{1}{\chi}}$ further increases to match the required increased in $\bar{d}_g$.

Appendix A.3 $p_g G/Y$ 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

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

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

The first term inside the parenthesis is increasing in $\bar{d}_g$ whenever $\epsilon \geq 1$, since we have shown that p_c and p_d decrease with it. We have also shown that p_g increases with $\bar{d}_g$. Then, to show that $\frac{p_g G}{Y}$ increases with $\bar{d}_g$, we only need to show that $\frac{\bar{d}_g}{d_p(\bar{d}_g, p_c, p_d(p_c))}$ increases with $\bar{d}_g$. This is true when $d_p(\bar{d}_g, p_c, p_d(p_c))$ decreases with $\bar{d}_g$. We have shown before that the direct effect of $\bar{d}_g$ is to decrease it when $\gamma > \beta(\chi - 1) + 1$. However, there are indirect effects through p_d and p_c. To the extent that the changes in $\bar{d}_g$ are small (which is what we have been studying), then the direct effect will dominate since the price effects are second order.

Note also that the above shows that not only $p_g G/Y$ increases but also G/Y. Then, since $D_g/Y = \kappa_g G/Y$, this implies that D_g/Y also increases with $\bar{d}_g$. This completes the proof.

Appendix B Quantitative analysis

Appendix B.1 Equilibrium conditions

The maximization problem of software producers is

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

FOC are

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

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

This implies

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

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

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

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

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:

$$\begin{aligned}
\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)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} \frac{1}{\chi-1} \times \cdots \\
&\qquad \left(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}}}\right) \\
(\bar{d}_p)^{\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)(\bar{d}_p)^{\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}} \underline{d}_p\right)^{\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} \frac{1}{\chi-1} \\
\underline{d}_p &= \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)(\underline{d}_p)^{\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} \times \cdots \\
&\qquad \left(\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\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) \underline{d}_p + N_g \bar{d}_p \\
\Pi_z(p_z) &= (p_z)^{\frac{\chi}{1+\beta(\chi-1)}} \frac{1+\beta(\chi-1)}{\chi-1} \\
Y_z &= N_z \frac{\Pi_z(p_z)}{p_z} \frac{\chi}{1+\beta(\chi-1)}
\end{aligned}$$

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

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

And the remaining market clearing conditions are

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

And the free entry conditions are

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

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

Appendix B.2 Calibration

We externally calibrate $\theta = 2, \rho = 0.03, \chi = 3$, 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 percent, 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 $\beta = 0.7$ and $\gamma = 2 + \beta(\chi - 1)$ which ensure that a symmetric BGP equilibrium exist, and then discuss how sensible our results are to changes in these parameters.

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 with respect to $\bar{d}_g$ the optimal levels of software production for those firms obtaining contracts, we obtain the partial equilibrium responses:

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

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 set 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 $\alpha = 0.82$.